IBM

Implementing IBM System Networking 10Gb Ethernet Switches

- Introduction to IBM System Networking RackSwitch hardware
- Sample network design and implementation
- Switch troubleshooting and maintenance

Sangam Racherla
Ionut Bunea
David A Gray
Sebastian Oglaza
Jose Rodriguez Ruibal

Redbooks

ibm.com/redbooks

International Technical Support Organization

Implementing IBM System Networking 10Gb Ethernet Switches

June 2012

SG24-7960-00

Note: Before using this information and the product it supports, read the information in “Notices” on page ix.

First Edition (June 2012)

This edition applies to IBM System Networking 10Gb Top-of-Rack, and Embedded Switches from the IBM System Networking portfolio of products.

© Copyright International Business Machines Corporation 2012. All rights reserved.
Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

Notices ix
Trademarks x

Preface xi
The team who wrote this book xi
Now you can become a published author, too! xiii
Comments welcome. xiv
Stay connected to IBM Redbooks xiv

Chapter 1. Introduction to IBM System Networking 10Gb Ethernet products. 1
1.1 Overview 2
1.1.1 Product information and features 2
1.1.2 Reference architecture 2
1.1.3 Initial configuration 2
1.1.4 IBM RackSwitch Implementation 3
1.1.5 Embedded switch implementation 3
1.1.6 Maintenance and troubleshooting. 3
1.2 IBM System Networking 10Gb RackSwitch information 4
1.2.1 High reliability and availability. 5
1.3 IBM System Networking RackSwitch G8052 6
1.3.1 Switch inclusions. 6
1.3.2 IBM System Networking RackSwitch G8052 features 7
1.3.3 Features and specifications 7
1.3.4 RackSwitch G8052 LED status details 11
1.3.5 More information. 12
1.4 IBM System Networking RackSwitch G8124 12
1.4.1 Switch inclusions. 13
1.4.2 IBM System Networking RackSwitch G8124 features 14
1.4.3 Features and specifications 15
1.4.4 IBM System Networking RackSwitch G8124 LED status details 18
1.4.5 More Information. 19
1.5 IBM System Networking RackSwitch G8264 20
1.5.1 Switch inclusions. 21
1.5.2 IBM System Networking RackSwitch G8264 benefits. 21
1.5.3 Features and specifications 22
1.5.4 IBM System Networking RackSwitch G8264 LED status details 26
1.5.5 More information. 27
1.6 IBM BladeCenter switches 27
1.6.1 IBM Virtual Fabric 10Gb Switch Module for IBM BladeCenter 27
1.6.2 IBM 1/10 Uplink Ethernet Switch Module for IBM BladeCenter 33
1.6.3 IBM BladeCenter H. 38
1.6.4 IBM BladeCenter HT. 43
1.7 Connectors, cables, and options. 48
1.7.1 More information. 49
1.8 Product interoperability 49

Chapter 2. IBM System Networking Switch 10Gb Ethernet switch features 51
2.1 Virtual Local Area Networks 52
2.1.1 VLANs overview 52

© Copyright IBM Corp. 2012. All rights reserved.

2.1.2 VLANs and Port VLAN ID numbers . . . 52
2.1.3 Protocol-based VLANs . . . 56
2.2 Spanning Tree Protocol . . . 58
2.2.1 Rapid Spanning Tree Protocol . . . 59
2.2.2 Per-VLAN Rapid Spanning Tree Protocol (PVRST) . . . 59
2.2.3 Multiple Spanning Tree Protocol . . . 59
2.3 IP routing . . . 60
2.3.1 Static routes . . . 61
2.3.2 Equal-Cost Multi-Path static routes . . . 61
2.3.3 Routing Information Protocol . . . 61
2.3.4 Open Shortest Path First . . . 62
2.3.5 Border Gateway Protocol . . . 66
2.4 IP multicast . . . 67
2.4.1 Internet Group Management Protocol . . . 67
2.4.2 Protocol Independent Multicast . . . 69
2.5 IPv6 . . . 71
2.5.1 IPv6 address format . . . 71
2.5.2 IPv6 address types . . . 72
2.5.3 IPv6 address auto-configuration . . . 73
2.5.4 Neighbor Discovery protocol . . . 73
2.5.5 IPv6 support . . . 74
2.6 Monitoring . . . 74
2.6.1 Port mirroring . . . 74
2.6.2 ACL-based mirroring . . . 75
2.6.3 sFlow . . . 75
2.6.4 Remote Monitoring (RMON) . . . 76
2.7 High availability . . . 77
2.7.1 Trunking . . . 77
2.7.2 Virtual Link Aggregation Groups . . . 79
2.7.3 Hot links . . . 79
2.7.4 Fast Uplink Convergence . . . 80
2.7.5 NIC teaming and Layer 2 failover . . . 80
2.7.6 Virtual Router Redundancy Protocol . . . 82
2.7.7 Active Multipath Protocol . . . 86
2.7.8 Stacking . . . 88
2.8 Security . . . 89
2.8.1 Private VLANs . . . 89
2.8.2 Securing administration . . . 90
2.8.3 Authentication and authorization protocols . . . 91
2.8.4 MAC address notification . . . 95
2.8.5 802.1x Port-based network access control . . . 95
2.8.6 Access control lists . . . 97
2.8.7 VLAN maps . . . 100
2.8.8 Storm-control filters . . . 100
2.9 Quality of Service . . . 100
2.9.1 QoS overview . . . 100
2.9.2 Using ACL filters . . . 101
2.9.3 Summary of ACL actions . . . 101
2.9.4 ACL metering and re-marking . . . 102
2.9.5 DiffServ Code Points . . . 102
2.9.6 QoS 802.1p . . . 105
2.9.7 Queuing and scheduling . . . 105

Chapter 3. Reference architectures 107
3.1 Overview of the reference architectures 108
3.2 Top-of-Rack architecture 108
3.2.1 Layer 1 architecture 109
3.2.2 Layer 2 architecture 111
3.2.3 Layer 3 architecture 113
3.3 IBM BladeCenter architecture 116
3.3.1 Layer 1 architecture 117
3.3.2 Layer 2 architecture 119
3.3.3 Layer 3 architecture 120
3.4 Final architecture 122

Chapter 4. Initial configuration: IBM System Networking 10Gb Ethernet switches . 123
4.1 Overview of the initial setup 124
4.2 Administration interfaces 124
4.2.1 Console, Telnet, and Secure Shell (SSH) 125
4.2.2 Browser-Based interface 125
4.3 First boot of the RackSwitch G8264 switch 127
4.3.1 Logging on to the switch 128
4.3.2 Global Configuration mode 130
4.3.3 Setup tool 131
4.4 First boot of the Virtual Fabric 10Gb Switch Module embedded switch 142
4.4.1 Basic options 143
4.4.2 Setting an IP address 144
4.4.3 Advanced options 145
4.4.4 Telnet access 146
4.4.5 Web access 147
4.4.6 Setting the date and time 148
4.4.7 Firmware upgrade from the AMM web interface 149
4.4.8 Working with users and passwords 149
4.5 IBM System Networking Element Manager 150
4.5.1 IBM System Networking Element Manager solution architecture 150
4.5.2 IBM System Networking Element Manager solution requirements 152

Chapter 5. IBM System Networking RackSwitch implementation 155
5.1 Layer 1 implementation 156
5.1.1 Network topology for Layer 1 configuration 156
5.1.2 Port settings configuration 157
5.2 Layer 2 163
5.2.1 VLANs 163
5.2.2 Ports and trunking 170
5.2.3 Spanning Tree Protocol 179
5.2.4 Quality of Service 183
5.3 Layer 3 186
5.3.1 Basic IPv4 configuration 186
5.3.2 Basic IPv6 configuration 192
5.3.3 Border Gateway Protocol 219
5.4 High availability 222
5.4.1 Virtual Router Redundancy Protocol 223
5.4.2 Layer 2 Failover 230
5.4.3 Trunking 236
5.4.4 Hot Links 237
5.5 More information 238

Chapter 6. IBM Virtual Fabric 10Gb Switch Module implementation 239
6.1 Purpose of this implementation 240
6.2 Stacking 240
6.2.1 Stacking overview 241
6.2.2 Stacking requirements 241
6.2.3 Stacking limitations 241
6.2.4 Stack membership 242
6.3 Layer 1 implementation 255
6.3.1 Network topology for Layer 1 configuration 256
6.3.2 Port settings configuration 256
6.4 Layer 2 implementation 260
6.4.1 VLANs 261
6.4.2 Ports and trunking 265
6.4.3 Spanning Tree Protocol 269
6.4.4 Quality of Service 271
6.5 High availability 272
6.5.1 Stacking 272
6.5.2 Layer 2 Failover 273
6.5.3 Trunking 273
6.5.4 Hot Links 273
6.5.5 VRRP 274
6.6 IPv4 and IPv6 end-to-end connectivity verification 281
6.7 More information 284

Chapter 7. Maintenance and troubleshooting 285
7.1 Configuration management 286
7.1.1 Configuration files 286
7.1.2 Configuration blocks 286
7.1.3 Managing configuration files 286
7.1.4 Factory defaults 289
7.1.5 Password recovery 290
7.2 Firmware management 290
7.2.1 Firmware files 290
7.2.2 Boot Management Menu 291
7.2.3 Loading the new firmware 292
7.2.4 Recovering from a failed firmware upgrade 293
7.3 Logging and reporting 295
7.3.1 System logs 295
7.3.2 SNMP 297
7.3.3 Remote Monitoring (RMON) 301
7.3.4 Management applications 302
7.4 Troubleshooting 306
7.4.1 LEDs 306
7.4.2 Basic troubleshooting procedure 306
7.4.3 Connectivity troubleshooting 307

Appendix A. Configuration files 309
AGG-1: Aggregation switch (RackSwitch G8264) 310
AGG-2: Aggregation switch (RackSwitch G8264) 315
ACC-1: Access switch (RackSwitch G8124) 320
ACC-2: Access switch (RackSwitch G8124) 325
ACC-3: Access switch (Virtual Fabric 10G Switch Module stack) 329

Related publications 333

Locating the web material 333
IBM Redbooks 333
Other publications 333
Online resources 336
Help from IBM 336

Index 337

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs.

© Copyright IBM Corp. 2012. All rights reserved.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at http://www.ibm.com/legal/copytrade.shtml

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

AIX®
BladeCenter®
BNT®
Global Technology Services®
IBM®
iDataPlex®
Netcool®
Power Systems™
Redbooks®
Redbooks (logo) ®
ServerProven®
System p®
System Storage®
System x®
Tivoli®
VMready®

The following terms are trademarks of other companies:

Intel, Intel logo, Intel Inside logo, and Intel Centrino logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Catalyst, the AMD Arrow logo, and combinations thereof, are trademarks of Advanced Micro Devices, Inc.

InfiniBand, and the InfiniBand design marks are trademarks and/or service marks of the InfiniBand Trade Association.

Novell, SUSE, the Novell logo, and the N logo are registered trademarks of Novell, Inc. in the United States and other countries.

QLogic, and the QLogic logo are registered trademarks of QLogic Corporation. SANblade is a registered trademark in the United States.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Other company, product, or service names may be trademarks or service marks of others.

Preface

In today's infrastructure, it is common to build networks based on 10 Gb Ethernet technology. The IBM® portfolio of 10 Gb systems networking products includes Top-of-Rack switches, and the embedded switches in the IBM BladeCenter® family. In 2010, IBM formed the IBM System Networking business (by acquiring BLADE Network Technologies), which is now focused on driving data center networking by using the latest Ethernet technologies.

The main focus of this IBM Redbooks® publication is on the IBM System Networking 10Gb Switch Modules, which include both embedded and Top-of-Rack (TOR) models. After reading this book, you can perform basic to advanced configurations of IBM System Networking 10Gb Switch Modules.

In this publication, we introduce the various 10 Gb switch models that are available today and then describe in detail the features that are applicable to these switches.

We then present two architectures that use these 10 Gb switches, which are used throughout this book. These designs are based on preferred practices and the experience of authors of this book. Our intention is to show the configuration of the different features that are available with IBM System Networking 10Gb Switch Modules. We follow the three-tier Data Center design, focusing on the Access and Aggregation Layers, because those layers are the layers that IBM System Networking Switches use.

We start our configuration with the initial setup of the switches, which is required to activate the switches. We also introduce the IBM System Networking Element Manager and its configuration.

In the IBM RackSwitch Implementation, we provide information and instructions for implementing 10 Gb Ethernet using the IBM Top-of-Rack switch models G8264R/F and G8124. Step by step instructions are included for implementing and configuring the most important functions of the IBM Networking Operating System (previously know a BLADEOS).

The Virtual Fabric 10Gb Switch Module implementation shows how a mixed environment of both stand-alone and embedded switches provide end-to-end communication in a data center, for servers that run different operating systems and IP protocol versions (IPv4 and IPv6).

Note: Because of the recent acquisition of IBM BNT® by IBM, some of the product names used in this publication might change.

The team who wrote this book

This book was produced by a team of specialists from around the world working at the International Technical Support Organization, San Jose.

© Copyright IBM Corp. 2012. All rights reserved.

Sangam Racherla is an IT Specialist and Project Leader working at the ITSO in San Jose, CA. He has 12 years of experience in the IT field and has been with the ITSO for the past eight years. Sangam has extensive experience in installing and supporting the ITSO lab equipment for various IBM Redbooks projects. He has expertise in working with Microsoft Windows, Linux, IBM AIX®, IBM System x®, and IBM System p® servers, and various SAN and storage products. Sangam holds a degree in electronics and communication engineering.

Ionut Bunea is a Certified IT Specialist working for IBM Global Technology Services® in Romania with more than 9 years of networking experience in designing implementations of wide area, campus, and data center network solutions. Ionut joined IBM in 2008 and is a Cisco Certified Networking Professional (CCNP), Cisco Certified Design Professional (CDDP), and Juniper Networks Certified Internet Specialist (JNCIS).

David A Gray is a Senior IT Specialist working for the System x and IBM System Storage® brands at IBM Australia in Brisbane, Queensland. He has over 27 years of professional experience in the IT industry, over 20 of them in the finance sector, and the last three of them with IBM. He instructs IBM Business Partners and customers about how to configure and install System x, BladeCenter, Systems Director, System Storage, and VMware. He is an IBM Certified Systems Expert - System x BladeCenter.

Sebastian Oglaza joined IBM Global Technology Services in 2006. Since then, he has been working as a Network Specialist in the Integrated Communications Services group. During this time, he participated in numerous projects in both design and implementation roles. He is an expert in data and voice networking. He holds a CCIE certification in Routing and Switching.

Jose Rodriguez Ruibal is a Technical Sales Leader for the IBM System x Networking team that covers the southwest Europe region. He has more than 15 years of experience in IT, and has worked for IBM for almost ten years. His experience includes serving as Benchmark Manager in the IBM PSSC Benchmark Center in Montpellier, working as an IT Architect for Nokia Siemens Network in Finland, and IT Architect and Team Leader for the IBM STG OEM and Next Generation Networks teams in EMEA. Before joining IBM, he worked for Red Hat and other consulting firms. He holds MSc and BSc degrees in Computer Engineering and Computer Systems from Nebrija University, Madrid. His areas of expertise include Business Development, Strategic OEM Alliances, long-term IT projects in the telecom, media, and defense industries, high-level IT architecture and complex solutions design, Linux, and all x86 hardware. Jose has co-authored other IBM Redbooks publications on networking products from Juniper, Linux solutions, IBM x86 servers, and Performance Tuning for x86 servers.

Figure 1 shows the team.

Figure 1 Jose, David, Sangam, Sebastian, and Ionut

Thanks to the following people for their contributions to this project:

Ann Lund, Jon Tate, David Watts
International Technical Support Organization, San Jose

Nghiem V. Chu, Kam-Yee (Johnny) Chung, Michael Easterly, David Faircloth, David Iles, Jeffery M. Jaurigui, Harry W. Lafnear, Lan T. Nguyen, Tuan A. Nguyen, Pushkar B. Patil, William V. (Bill) Rogers, Rakesh Saha, Hector Sanchez, Tim Shaughnessy, Selvaraj Venkatesan
IBM System Networking Team, San Jose

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an ITSO residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

We want our books to be as helpful as possible. Send us your comments about this book or other IBM Redbooks publications in one of the following ways:

- Use the online **Contact us** review Redbooks form found at:

  ibm.com/redbooks

- Send your comments in an email to:

  redbooks@us.ibm.com

- Mail your comments to:

  IBM Corporation, International Technical Support Organization
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400

Stay connected to IBM Redbooks

- Find us on Facebook:

  http://www.facebook.com/IBMRedbooks

- Follow us on Twitter:

  http://twitter.com/ibmredbooks

- Look for us on LinkedIn:

  http://www.linkedin.com/groups?home=&gid=2130806

- Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

  https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm

- Stay current on recent Redbooks publications with RSS Feeds:

  http://www.redbooks.ibm.com/rss.html

1

Introduction to IBM System Networking 10Gb Ethernet products

Networks are changing. Voice, video, storage, and data are quickly converging onto a single backbone. Growth in cloud services and Web 2.0 multimedia content is pushing bandwidth demand to the edge. These bandwidth demands are also increasing as clients employ virtualization and focus on maximizing server usage. The next level of network consolidation has to do with I/O and storage.

Gigabit Ethernet link aggregation is reaching its limits because of cable management issues, power and cooling costs for multiple switches, and administrative impact and limited bandwidth.

Multicore processing environments with large memory configuration require substantial bandwidth. As the demands for mission-critical and real-time applications continue to expand, servers must offer higher processing capabilities to keep up. Because existing switch architectures cannot handle the data throughput required for these applications, they are becoming bottlenecks.

This chapter provides details about the IBM System Networking 10Gb switches that are used for the two implementations in this publication. We first describe the architectures in Chapter 3, “Reference architectures” on page 107, and the actual configuration procedures are documented in Chapter 5, “IBM System Networking RackSwitch implementation” on page 155 and Chapter 6, “IBM Virtual Fabric 10Gb Switch Module implementation” on page 239.

Although there are many possible implementations of 10 Gb infrastructures, we focus primarily on the IBM System Networking 10Gb switches and the specific features that are available and are related to our reference architecture.

Terminology: Because of the recent acquisition of BNT by IBM, some of the product names used in this publication might change.

© Copyright IBM Corp. 2012. All rights reserved.

1.1 Overview

In today's infrastructure, it is common to build networks based on 10 Gb Ethernet technology. The IBM portfolio of 10 Gb systems networking products includes Top-of-Rack switches, and the embedded switches in the IBM BladeCenter family. In 2010, IBM formed the IBM System Networking business (by acquiring BLADE Network Technologies), which is now focused on driving data center networking by using the latest in Ethernet technologies.

The main focus of this publication is the IBM System Networking 10Gb switches, which include both embedded and Top-of-Rack (TOR) models. After reading this book, you can perform basic to advanced configurations of IBM System Networking 10Gb switches.

1.1.1 Product information and features

The IBM System Networking portfolio of products includes 10 Gb switch models in both embedded and Top-of-Rack (TOR) types. In this chapter, we introduce the various switch models that are available today and then describe in detail the features that are applicable to these switches in Chapter 2, "IBM System Networking Switch 10Gb Ethernet switch features" on page 51.

1.1.2 Reference architecture

In Chapter 3, "Reference architectures" on page 107, we present two architectures that are used throughout this book. The designs are based on preferred practices and the experience of the authors of this book. Our intention is to show the configuration of the different features available with the switches. We follow the three-tier Data Center design, focusing on the access and aggregation layers, because those layers are the layers IBM System Networking products use.

Both designs differ in how the Access Layer is implemented:

- The Top-of-Rack architecture uses a pair of stand-alone switches as the access layer.
- The BladeCenter architecture uses a pair of Virtual Fabric 10Gb Switch Modules in the IBM BladeCenter chassis as the access layer.
- The architectures share a common aggregation layer.

1.1.3 Initial configuration

In Chapter 4, "Initial configuration: IBM System Networking 10Gb Ethernet switches" on page 123, we describe the initial configuration steps you must perform to activate the TOR and embedded switches.

This chapter also provides basic introduction to the architecture of IBM System Networking Element Manager (SNEM) and then provides links to the appropriate documentation.

The steps in this chapter cover the following elements for the hardware:

- Terminal connection
- Setting up the IP address of the switch
- Configuring date and time
- Security

1.1.4 IBM RackSwitch Implementation

Chapter 5, “IBM System Networking RackSwitch implementation” on page 155 provides information and instructions for implementing the 10 Gb Ethernet with Top-of-Rack switch models G8264R/F and G8124. As described in the reference architecture in Chapter 3, “Reference architectures” on page 107, we present a step by step guide for implementing and configuring the most important functions of the IBM Networking OS.

This chapter covers implementation aspects that pertain to OSI Layer 1 - 3, as follows:

- Layer 1: Configuration, information, and statistics commands related to Layer 1 operation. Port configuration in terms of speed and duplex, link status, errors, and so on.
- Layer 2: Configuration, information, and statistics commands related to Layer 2 operation. VLANs, ports and trunking, Spanning Tree Protocol, QoS, and high availability mechanisms.
- Layer 3: Configuration, information, and statistics commands related to Layer 3 operation. Basic IP routing, dynamic routing protocols, high availability mechanisms (Virtual Router Redundancy Protocol (VRRP)), and IPv6.

1.1.5 Embedded switch implementation

Chapter 6, “IBM Virtual Fabric 10Gb Switch Module implementation” on page 239 shows how a mixed environment of both stand-alone and embedded switches provide end-to-end communication in a data center, for servers that run different operating systems and IP protocol versions (IPv4 and IPv6).

The topics described in this chapter include:

- Stacking implementation: Overview, requirements, limitations, configuration, operation, and redundancy
- Layer 1 implementation: Ports connection, configuration, and verification
- Layer 2 implementation: VLANs, tagging, trunking, Spanning Tree Protocol, and Quality of Service (QoS)
- High availability: Stacking, Layer 2 failover, trunking, Hot Links, and VRRP at the aggregation layer

1.1.6 Maintenance and troubleshooting

Chapter 7, “Maintenance and troubleshooting” on page 285 describes some useful elements that can help you with the maintenance and troubleshooting of IBM System Networking 10Gb switches. Some of the topics described there are:

- Managing configuration files
- Factory defaults
- Firmware management
- Logging and reporting
- Basic troubleshooting procedures.

In Appendix A, “Configuration files” on page 309, we conclude this publication with the final working configurations of all the equipment used for the implementations described in the reference architectures.

1.2 IBM System Networking 10Gb RackSwitch information

In this section, we provide detailed information about the IBM System Networking 10Gb Switches summarized in Table 1-1.

Table 1-1 IBM System Networking 10Gb products covered in this publication

Description	Switch Model
IBM Top-of-Rack switches	
	IBM System Networking RackSwitch G8052
	IBM System Networking RackSwitch G8124
	IBM System Networking RackSwitch G8264
IBM BladeCenter switches	
	IBM BladeCenter H Chassis
	IBM BladeCenter HT Chassis
	IBM System Networking Virtual Fabric 10Gb Network Switch
	IBM System Networking 1/10Gb Uplink Ethernet Switch Module for IBM BladeCenter

We also provide some information about the IBM System Networking 10Gb options, as shown in Table 1-2.

Table 1-2 Networking cables, connectors, and miscellaneous options

Description	IBM Part No.	Feature Code[a]
1 Gb options		
IBM SFP 1000Base-T Transceiver	81Y1618	EB29
IBM SFP 1000Base-SX SR Fiber Transceiver	81Y1622	EB2A
IBM SFP 1000Base-LX LR Fiber Transceiver	90Y9424	
IBM SFP 1000Base-ZX Fiber Transceiver	90Y9418	
0.6 m Blue Cat5e Cable	40K5679	
1.5 m Blue Cat5e Cable	40K8785	
3 m Blue Cat5e Cable	40K5581	
10 m Blue Cat5e Cable	40K8927	
25 m Blue Cat5e Cable	40K8930	
10 Gb options		
IBM SFP+ SR Transceiver	46C3447	EB28
IBM SFP+ LR Transceiver	90Y9412	
IBM SFP+ ER Transceiver	90Y9415	

1 m IBM Passive DAC SFP+ Cable	90Y9427	
3 m IBM Passive DAC SFP+ Cable	90Y9430	
5 m IBM Passive DAC SFP+ Cable	90Y9433	
8.5 m IBM Passive DAC SFP+ Cable	90Y9436	
40 Gb options		
IBM QSFP+ SR Transceiver	49Y7884	EB27
1 m IBM QSFP+ DAC Break Out Cable	49Y7886	EB24
3 m IBM QSFP+ DAC Break Out Cable	49Y7887	EB25
5 m IBM QSFP+ DAC Break Out Cable	49Y7888	EB26
1 m IBM QSFP+-to-QSFP+ Cable	49Y7890	
3 m IBM QSFP+-to-QSFP+ Cable	49Y7891	
10 m IBM QSFP+ MTP Optical Cable	90Y3519	
30 m IBM QSPF+ MTP Optical Cable	90Y3521	
Miscellaneous 0ptions		
IBM 19" Flexible 4 Post Rail Kit	49Y4284	
1 m LC-LC Fiber Cable (networking) - Optical	88Y6851	
5 m LC-LC Fiber Cable (networking)- Optical	88Y6854	
25 m LC-LC Fiber Cable (networking) - Optical	88Y6857	

a. A four-digit feature code is used to specify a component that is configurable in the IBM configurators, the eBusiness Configurator (eConfig), and the IBM Web-based Hardware Configurator, used in US and Canada. Options that do not have feature codes cannot be integrated by IBM into a system using these IBM configurators. To purchase options separately, refer to the option part number provided in this table.

Although this list is not a complete list of all available components, these components are the most relevant to the implementations that are referenced in this publication.

1.2.1 High reliability and availability

Because the IBM System Networking 10 Gb RackSwitches offer integrated, high-availability support in both Layer 2 and 3, this support minimizes single points of failure, ensuring network reliability and performance.

Layer 2 - high availability is supported with Link Aggregation Control Protocol (LACP), Rapid Spanning Tree, Cisco UplinkFast compatibility, PortFast compatibility, 802.1Q VLANs, broadcast storm control, and controlled link failover with NIC teaming. VRRP Hot-Standby further enables the effective use of Layer 2 NIC teaming failover.

Layer 3 - high availability is supported in a special extended version of VRRP that allows multiple 10 Gb switches to process traffic in an active-active configuration and concurrently process traffic (not sit in standby). This configuration enables maximum performance, and allows easy failover in the unlikely event of a system problem.

These features are covered in more detail in Chapter 2, “IBM System Networking Switch 10Gb Ethernet switch features” on page 51.

1.3 IBM System Networking RackSwitch G8052

This switch is a Top-of-Rack switch designed for a data center. It combines great performance, server-like airflow for cooling, and low-power consumption in a virtualization-ready package. Figure 1-1 shows the IBM System Networking RackSwitch G8052 Top-of-Rack (TOR) Switch.

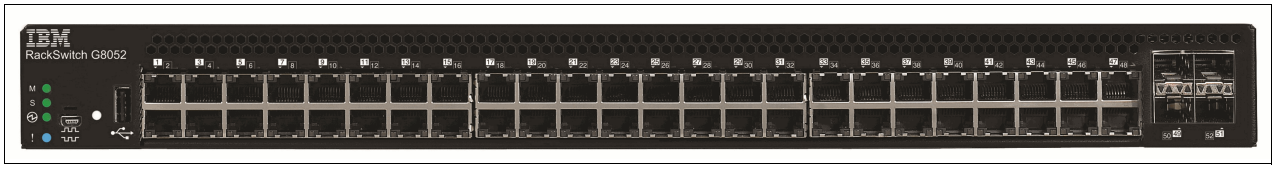

Figure 1-1 IBM System Networking RackSwitch G8052 TOR Switch

The RackSwitch G8052 is a Top-of-Rack data center switch that delivers unmatched line-rate Layer 2/3 performance at an attractive price. It has 48 10/100/1000 BASE-T RJ45 ports and four 10 Gigabit Ethernet SFP+ ports, and includes hot-swap redundant power supplies and fans standard, minimizing your configuration requirements. Unlike most rack equipment that cools from side to side, the RackSwitch G8052 has rear-to-front or front-to-rear airflow that matches server airflow.

For 10 Gb uplinks, there is a choice of either SFP+ transceivers (SR or LR) for longer distances or more cost-effective and lower-power-consuming options such as SFP+ direct-attached cables (DAC or Twinax cables), which can be 1 - 7 meters in length and are ideal for connecting to another Top-of-Rack switch, or even connecting to an adjacent rack. These options are covered in more detail in 1.7, “Connectors, cables, and options” on page 48.

This switch is ideal for connecting servers with 1 Gb Network interfaces into a 10 Gb network while maintaining high port density.

Table 1-3 show the part numbers used to order the IBM System Networking RackSwitch G8052.

Table 1-3 IBM System Networking RackSwitch G8052 part numbers

Description	IBM Part No.	IBM Power Systems™ MTM/FC
IBM System Networking RackSwitch G8052R (Rear-to-Front)	7309G52	1455-48E
IBM System Networking RackSwitch G8052F (Front-to-Rear)	730952F	

1.3.1 Switch inclusions

The module part numbers include the following items:

- One RackSwitch G8052R / G8052F
- Generic Rack Mount Kit (2-post)

- Mini-USB to DB9 serial cable (3 m)
- Comes with an IBM limited 3-year hardware warranty and includes a 3-year software license, providing entitlement to upgrades over that period
- Two power cords, depending on the country of purchase (Make sure that you include these cords in your configuration.)

Transceivers: Small form-factor pluggable plus (SFP+) transceivers are not included in the purchase of the switch. All 1/10 Gb transceivers require LC-to-LC cables.

1.3.2 IBM System Networking RackSwitch G8052 features

The RackSwitch G8052 provides the following features:

- High performance: The RackSwitch G8052 provides up to 176 Gbps throughput and supports four SFP+ 10 Gb uplink ports for a low oversubscription ratio, in addition to a low latency of 1.7 ms.
- Lower power and better cooling: The RackSwitch G8052 typically consumes just 120 W of power, a fraction of the power consumption of most competitive offerings. Unlike side-cooled switches, which can cause heat recirculation and reliability concerns, the G8052 rear-to-front or front-to-rear cooling design reduces data center air conditioning costs by matching airflow to the server's configuration in the rack. Variable speed fans assist in automatically reducing power consumption.
- Layer 3 Functionality: The G8052 switch includes Layer 3 functionality, which provides security and performance benefits, as inter-VLAN traffic can be processed at the access layer. This switch also provides the full range of Layer 3 static and dynamic routing protocols, including Open Shortest Path First (OSPF) and Border Gateway Protocol (BGP) for enterprise customers at no additional cost.
- VM-Aware Network Virtualization: IBM VMready® software on the switch simplifies configuration and improves security in virtualized environments. VMready automatically detects VM movement between physical servers and instantly reconfigures each VM's network policies across VLANs to keep the network up and running without interrupting traffic or impacting performance. VMready works with all leading hypervisors, such as VMware, Citrix Xen, KVM, and Microsoft.
- Fault tolerance: These switches learn alternative routes automatically and perform faster convergence in the unlikely case of a link, switch, or power failure. The switch uses proven technologies such as L2 trunk failover, advanced VLAN-based failover, VRRP, Hot Link, Uplink Failure Detection (UFD), IGMP v3 Snooping, and OSPF.
- Seamless interoperability: IBM RackSwitch switches interoperate seamlessly with other vendors' upstream switches.

1.3.3 Features and specifications

In this section, we list some of the hardware and software features and specifications of the RackSwitch G8052. For more details about these features, see Chapter 2, "IBM System Networking Switch 10Gb Ethernet switch features" on page 51.

Performance

The performance features and specifications of the RackSwitch G8052 are:

- Single switch ASIC design
- Full line rate performance
- 176 Gbps (full duplex) switching architecture
- Low latency: 1.7 ms

Hardware features

The hardware features of the RackSwitch G8052 are:

- Models:
  - RackSwitch G8052F (for front-to-rear cooling). The ports at the front of the rack match the airflow of IBM RackSwitch iDataPlex®.
  - RackSwitch G8052R (for rear-to-front cooling). The ports at the rear of the rack match the IBM System x and BladeCenter designs.
- Interface options:
  - Forty-eight 10/100/1000BaseT ports (RJ-45).
  - Four 10 GbE SFP+ ports.
  - One USB port for external storage devices[1].
  - An RS-232 serial console port that provides an additional means to install software and configure the switch module. This USB-style connector enables connection of a special serial cable that is supplied with the switch module.

Figure 1-2 shows the front view of the RackSwitch G8052 with the different ports.

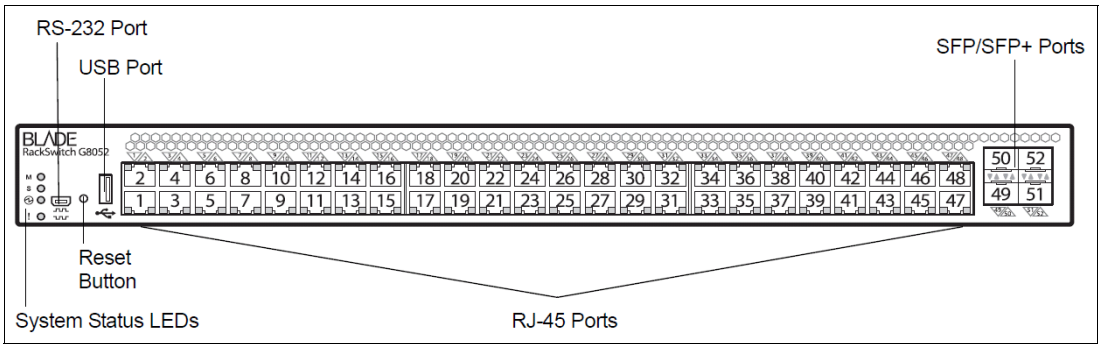

Figure 1-2 RackSwitch G8052 port types and locations (front view)

[1] If a USB drive is inserted into the USB port, you can copy files from the switch to the USB drive, or from the USB drive to the switch. You also can boot the switch by using software or configuration files found on the USB drive.

Figure 1-3 shows the rear view of the switch.

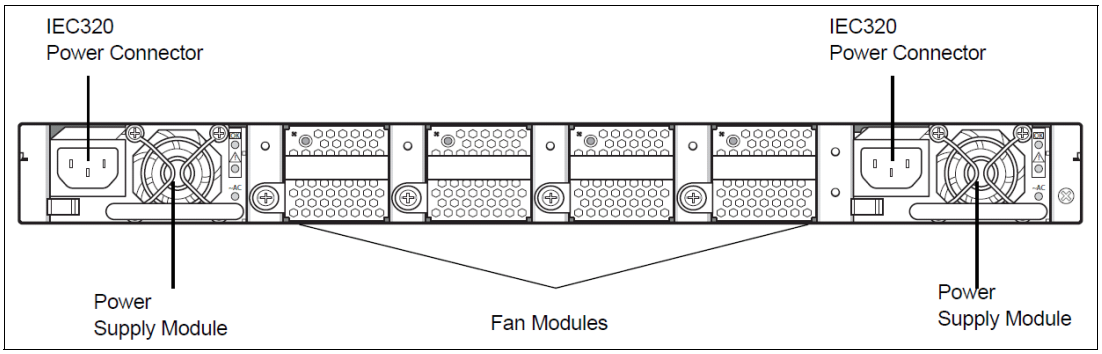

Figure 1-3 RackSwitch G8052 connections and modules (rear view)

Software features

The software features for the RackSwitch G8052 are:

- Security:
  - LDAP
  - 802.1x with VLAN assignment
  - Private VLAN edge
  - RADIUS
  - TACACS+
  - Wirespeed filtering
  - Flexible ACL combinations – L2-L4 criteria: Source and destination MAC, IP, and TCP/UDP ports
  - SSH v1 and v2
  - HTTPS Secure BBI
  - MAC address move notification
  - SCP
  - Shift B Boot menu (password recovery/factory default)
- VLANs:
  - Port-based VLAN
  - 4096 VLAN IDs supported
  - 1024 Active VLANs (802.1Q)
  - 802.1x with Dynamic VLAN assignment
  - Private VLAN Edge
- Trunking:
  - LACP
  - Static trunks (EtherChannel)
  - Configurable trunk hash algorithm
- Spanning Tree:
  - Multiple Spanning Tree (802.1 s)
  - Rapid Spanning Tree (802.1 w)

  - Fast uplink convergence
  - PVRST+
- Quality of service:
  - QoS 802.1p
  - DSCP
  - Weighted round robin
  - Metering
  - 4 MB buffers for queuing
- Routing protocols:
  - 128 static routes
  - Layer 2/3 static routes
  - RIP v1 and v2
  - OSPF v3
  - BGP
  - IPv6
- High availability:
  - Uplink failure detection
  - Hot links
  - VRRP support
  - Active multipath
  - Layer 2 failover
- Multicast: IGMP v1, v2, and v3 snooping with 2 K IGMP groups
- Monitoring:
  - Port mirroring
  - ACL-based mirroring
  - sFlow Version 5
- Virtualization: VMready VI API support

Management features

RackSwitch G8052 supports the following management clients:

- SNEM
- Industry Standard Command Line Interface (ISCLI)
- Browser-based client, SSH, or Telnet
- Netboot

Standard protocols

The RackSwitch G8052 supports the following standard protocols:

- SNMP v1, v2c, and v3
- RMON
- Secondary NTP Support
- Accept DHCP
- LLDP
- 32K MAC Table

- 9 K Jumbo Frames
- 802.3X Flow Control

1.3.4 RackSwitch G8052 LED status details

Figure 1-4 shows the LED indicators as they appear on the switch. Their meanings are explained in Figure 1-5.

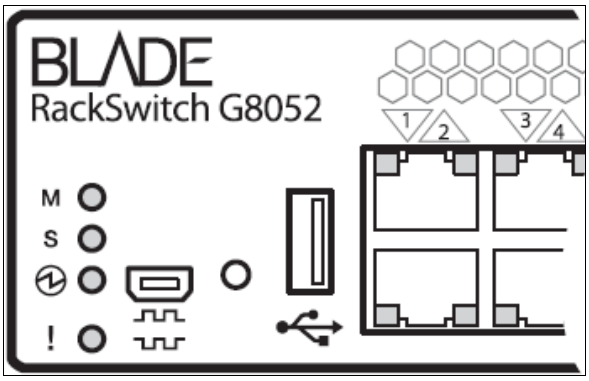

Figure 1-4 The location of LEDs on the RackSwitch G8052

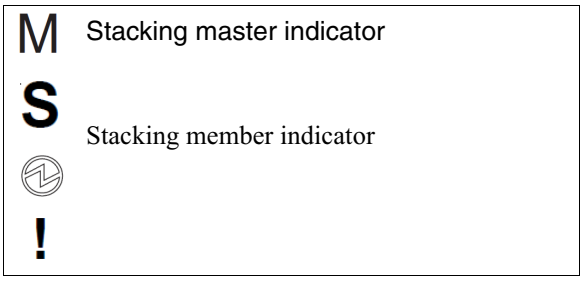

Figure 1-5 Indicator LEDs and their meanings

Table 1-4 shows the different System LED statuses for the RackSwitch G8052.

Table 1-4 RackSwitch G8052 System LED Status

Function	Stacking master indicator[a]	Stacking member indicator[a]	Power supply	Service
Total Power Failure	Off	Off	Off	Off
Service Required	Flash green	Flash green	Flash green	Flash Blue
Power Supplies OK	N/A	N/A	Solid green	N/A
Power Supply Failure	N/A	N/A	Flash green	N/A
Fans OK	N/A	N/A	N/A	N/A
Fan Failure	N/A	N/A	N/A	N/A
Stack Master	On	Off	N/A	Off
Stack Backup/Member	Off	On	N/A	On
Stack Error	On	On	N/A	On

Function	Stacking master indicator[a]	Stacking member indicator[a]	Power supply	Service
Non-Stack Member	Off	Off	N/A	Off

a. Stacking for the RackSwitch G8052 is not currently supported, but these indicators remain for possible future feature releases.

1.3.5 More information

For more about the RackSwitch G8052 and the LED status information, see the following resources:

- IBM System Networking RackSwitch G8052 Announcement Letter:

  http://www.ibm.com/common/ssi/cgi-bin/ssialias?infotype=AN&subtype=CA&appname=gpateam&supplier=872&letternum=ENUSAG11-0005&pdf=yes
- *IBM System Networking RackSwitch G8052*, TIPS0813
- *IBM System Networking RackSwitch G8052 Installation Guide*:

  http://www-01.ibm.com/support/docview.wss?uid=isg3T7000287&aid=1
- *IBM System Networking RackSwitch G8052 Application Guide (6.6)*:

  http://www.ibm.com/support/docview.wss?uid=isg3T7000353
- *IBM System Networking RackSwitch G8052 Browser-Based Interface Quick Guide*:

  http://www.ibm.com/support/docview.wss?uid=isg3T7000348
- *IBM System Networking RackSwitch G8052 ISCLI Command Reference*:

  http://www.ibm.com/support/docview.wss?uid=isg3T7000344
- *IBM System Networking RackSwitch G8052 Menu-Based CLI Reference Guide*:

  http://www.ibm.com/support/docview.wss?uid=isg3T7000347

1.4 IBM System Networking RackSwitch G8124

The IBM System Networking RackSwitch G8124 is designed with top performance in mind. This low-latency switch provides line-rate, high-bandwidth switching, filtering, and traffic queuing without delaying data.

Figure 1-6 shows the RackSwitch G8124 TOR switch.

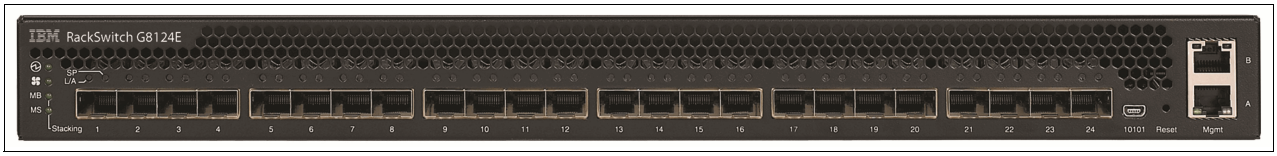

Figure 1-6 IBM System Networking RackSwitch G8124 TOR switch

With support for 1G or 10G, this switch is designed for those clients that are using 10G Ethernet today or plan to do so in the future. This switch was the first Top-of-Rack 10Gb switch for IBM System x designed to support IBM Virtual Fabric, which helps clients reduce cost and complexity when it comes to the I/O requirements of most virtualization deployments today. Virtual Fabric can help clients reduce the number of multiple I/O adapters down to a single dual-port 10G adapter, in addition to reducing the number of cables and upstream switch ports required. Virtual Fabric allows clients to carve up a dual-port 10G adapter into eight virtual NICs (vNICs) and create dedicated virtual pipes between the adapter and the switch for optimal performance, plus higher availability and better security. This functionality provides the ability to dynamically allocate bandwidth per vNIC in increments of 100 Mb, while being able to adjust over time without downtime.

With the flexibility of the G8124 switch, clients can take advantage of the technologies that they require for multiple environments. For 10 Gb uplinks, they have a choice of either SFP+ transceivers (SR or LR) for longer distances, or more cost-effective and lower-power-consuming options, such as SFP+ direct-attached cables (DAC or Twinax cables), which can be 1 - 7 meters in length and are ideal for connecting chassis together, connecting to a Top-of-Rack switch, or even connecting to an adjacent rack. These technologies are covered in more detail in 1.7, “Connectors, cables, and options” on page 48.

Table 1-5 shows the part numbers used to order the IBM System Networking RackSwitch G8124.

Table 1-5 IBM System Networking RackSwitch G8124 part numbers

Description	IBM Part No.	Power Systems MTM/FC
IBM System Networking RackSwitch G8124R (Rear-to-Front)	0446017	
IBM System Networking RackSwitch G8124F (Front-to-Rear)	7309BF9	
IBM System Networking G8124DC (DC Power)	7309BD5	
Enhanced models		
IBM System Networking RackSwitch G8124E-R (Rear-to-Front)	7309BR6	1455-24E
IBM System Networking RackSwitch G8124E-F (Front-to-Rear)	7309BF7	

1.4.1 Switch inclusions

The RackSwitch G8124 module part numbers include the following items:

- One IBM System Networking RackSwitch G8124
- Generic Rail Mount Kit (2-post)
- Mini-USB to DB9 serial cable (3 m)
- Publication Group
- An IBM limited 3-year hardware warranty

Transceivers: Small form-factor pluggable plus (SFP+) transceivers are not included in the purchase of the switch. All 1/10 Gb transceivers require LC-to-LC cables.

1.4.2 IBM System Networking RackSwitch G8124 features

The RackSwitch G8124 offers the following feature benefits:

- High performance: The 10G Low Latency (<700 ns) switch provides the best combination of low latency, non-blocking line-rate switching and ease of management.
- Lower power and better cooling: The RackSwitch G8124 uses as little power as two 60 W light bulbs, which is a fraction of the power consumption of most competitive offerings. Unlike side-cooled switches, which can cause heat recirculation and reliability concerns, the G8124 rear-to-front cooling design reduces data center air conditioning costs by having airflow match the servers in the rack. In addition, variable speed fans assist in automatically reducing power consumption.
- Virtual Fabric: Virtual Fabric can help customers address I/O requirements for multiple NICs while also helping reduce cost and complexity. Virtual Fabric for IBM allows the carving up of a physical NIC into multiple virtual NICs (2 - 8 vNICs) and creates a virtual pipe between the adapter and the switch for improved performance, availability, and security while reducing cost and complexity.
- VM-aware networking: VMready software on the switch helps reduce configuration complexity while improving security levels in virtualized environments. VMready automatically detects virtual machine movement from one physical server to another, and instantly reconfigures each VM's network policies across VLANs to keep the network running without interrupting traffic or impacting performance. VMready works with all leading VM providers, such as VMware, Citrix, Xen, and Microsoft.
- Layer 3 functionality: The switch includes Layer 3 functionality, which provides security and performance benefits as inter-VLAN traffic stays within the chassis. This switch also provides the full range of Layer 3 protocols from static routes for technologies such as Open Shortest Path First (OSPF) and Border Gateway Protocol (BGP) for enterprise customers.
- Active MultiPath (AMP): Effectively doubles bandwidth by allowing all uplink ports to be active/active, eliminating cross-stack traffic, and providing up to 900 Gbps aggregate bandwidth between servers. Built-in fault tolerance constant health checking ensures maximum availability.
- Seamless interoperability: iBM RackSwitch switches interoperate seamlessly with other vendors' upstream switches. For more information, see the following note box.

  Tolly Reports: Tolly Functionality and Certification: RackSwitch G8000 and G8124 and Cisco Catalyst Interoperability Evaluation.

  http://tolly.com/ts/2009/blade/G8100/Tolly209116BladeRackSwitchInteroperability.pdf

- Fault tolerance: These switches learn alternative routes automatically and perform faster convergence in the unlikely case of a link, switch, or power failure. The switch uses proven technologies, such as L2 trunk failover, advanced VLAN-based failover, VRRP, Hot Link, Uplink Failure Detection (UFD), IGMP V3 snooping, and OSPF.
- Converged fabric: The switch is designed to support Converged Enhanced Ethernet (CEE) and connectivity to Fibre Channel over Ethernet (FCoE) gateways. CEE helps enable clients to combine storage, messaging traffic, VoIP, video, and other data on a common data center Ethernet infrastructure. FCoE helps enable highly efficient block storage over Ethernet for consolidating server network connectivity. As a result, clients can deploy a single server interface for multiple data types, which can simplify both deployment and management of server network connectivity, while maintaining the high availability and robustness required for storage transactions.

1.4.3 Features and specifications

In this section, we list some of the hardware and software features and specifications of the RackSwitch G8124. For more details about these features, see Chapter 2, “IBM System Networking Switch 10Gb Ethernet switch features” on page 51.

Performance

The performance specifications are:

- 100% line rate performance
- Latency under 700 ns
- 480 Gbps non-blocking switching throughput (full duplex)

Hardware features

Here are the hardware models available for the RackSwitch G8124:

- Standard models:
  - RackSwitch G8124R (Rear-to-Front) IBM PN 0446017: The Rear-to-Front airflow is ideal for servers or blade chassis with ports in the back of the rack.
  - RackSwitch G8124F (Front-to-Rear) IBM PN 7309BF9: The Front-to-Rear airflow is ideal for IBM iDataplex or when you need ports in the front of the rack.
  - RackSwitch G8124DC (DC Power) IBM PN 7309CD9: The rear-to-Front airflow is ideal for servers or blade chassis with ports in the back of the rack.
- Enhanced models:
  - RackSwitch G8124E-R (Rear-to-Front) IBM PN 7309BR6: The Rear-to-Front airflow is ideal for servers or blade chassis with ports in the back of the rack.
  - RackSwitch G8124E-F (Front-to-Rear) IBM PN 7309BF7: The Front-to-Rear airflow is ideal for IBM iDataplex or when you need ports in the front of the rack.

Here are the hardware features for the RackSwitch G8124:

- Interface options:
  - Twenty-four 10G SFP+ fiber connectors.
  - 2x 10/100/1000 Ethernet RJ45 ports for management.
  - One mini-USB Console port for management that provides an additional means to install software and configure the switch module. This USB-style connector enables connection of a special serial cable that is supplied with the switch module.
  - Server-like port orientations, enabling short and simple cabling.
  - Active DAC support for interoperability with Cisco Nexus 5k and Brocade.
- LEDs:
  - System LEDs to indicate status.
  - LEDs to indicate master/member[2].
- Airflow:
  - Rear-to-front cooling.
  - Redundant variable speed fans for reduced power draw.

[2] Reserved for future OS releases

- Power:
  - The AC-Powered G8124 has dual load-sharing internal power modules, with 50 - 60 Hz and 100 - 240 VAC auto-switching per module.
  - The nominal power for the G8124 ranges from 115 W to 168 W depending on the speed of the port (1G/10G), type of transceivers (SR or DAC), and number of active ports.
- Mean time between failures (MTBF): 189,060 hrs with ambient operating temperature of 40 ºC. MTBF is calculated by using the Telcordia Technologies Reliability Prediction Procedure for Electronic Equipment (SR-332 issue 2), Parts Count (method 1 case 1) failure rate data.

Figure 1-7 shows front view of the RackSwitch G8124 with the different ports that are described previously. Figure 1-8 shows the rear view of the switch for both the AC and DC models.

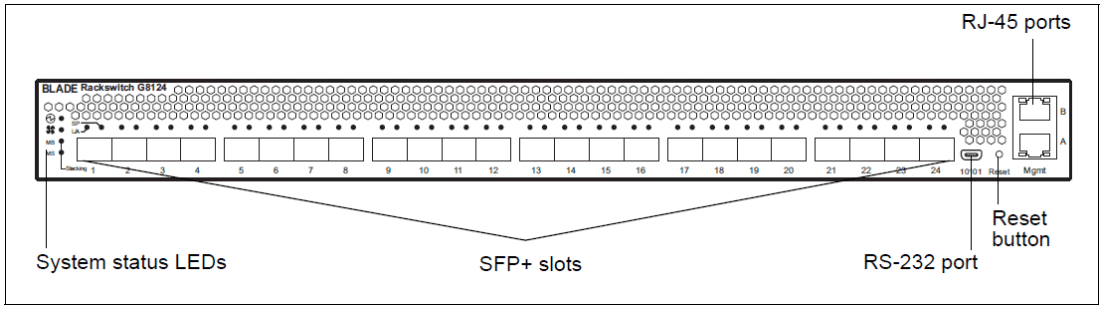

Figure 1-7 RackSwitch G8124 front panel

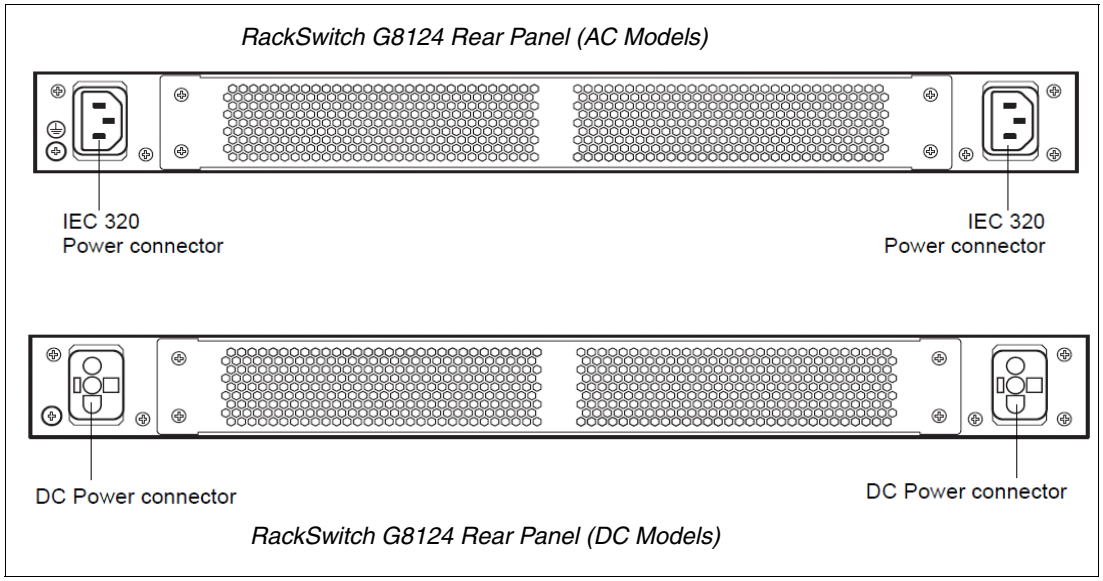

Figure 1-8 RackSwitch G8124 rear panel (AC and DC models)

Software features

The software features for the RackSwitch G8124 are:

- Security:
  - RADIUS
  - TACACS+

  - SCP
  - Wirespeed filtering: Allow and deny
  - SSH v1 and v2
  - HTTPS Secure BBI
  - Secure interface login and password
  - MAC address move notification
  - Shift B Boot menu (password recovery/factory default)
- VLANs:
  - Port-based VLANs
  - 4096 VLAN IDs supported
  - 1 k VLANs (802.1Q)
  - Private VLAN Edge
- FCoE/Lossless Ethernet:
  - 802.1 Data Center Bridging
  - Priority-based Flow Control (PFC)
  - Enhanced Transmission Selection (ETS)
  - Data Center Bridge Exchange protocol (DCBX)
  - FIP Snooping
  - Fibre Channel over Ethernet (FCoE)
  - Converged Enhanced Ethernet (CEE)
- Trunking:
  - LACP
  - Static Trunks (EtherChannel)
  - Configurable Trunk Hash algorithm
- Spanning Tree:
  - Multiple Spanning Trees (802.1 s)
  - Rapid Spanning Tree (802.1w)
  - PVRST+
  - Fast Uplink Convergence
  - BPDU guard
- Quality of Service:
  - QoS 802.1p (priority queues)
  - DSCP remarking
  - Metering
- Routing protocols:
  - RIP v1/v2
  - OSPF
  - BGP
- High availability:
  - Uplink failure detection
  - Hot Links
  - VRRP
  - Active MultiPath (AMP)
- Multicast:
  - IGMP Snoopingv1, v2, and v3 with 2 K IGMP groups
  - Protocol Independent Multicast (PIM sparse mode/dense mode)

- Monitoring:
  - Port mirroring
  - ACL-based mirroring
  - sFlow Version 5
- Virtualization:
  - VMready with VI API support
  - vNIC MIB support for SNMP Management features
  - Netboot
- Upgrades:
  - Upgrade firmware through serial or TFTP
  - Dual software images

Management features

RackSwitch G8124 supports the following management clients:

- IBM System Networking Element Manager
- isCLI (Cisco-like)
- Scriptable CLI
- Browser-based client or Telnet

Standard protocols

RackSwitch G8124 supports the following standard protocols:

- IPv6
- SNMP v1, v2c, and v3
- RMON
- Secondary NTP support
- Accept DHCP
- DHCP Relay
- LLDP
- 16 K MAC table
- 9 K jumbo frames
- 802.3X flow control

1.4.4 IBM System Networking RackSwitch G8124 LED status details

Figure 1-9 shows the LED indicators as they appear on the switch.

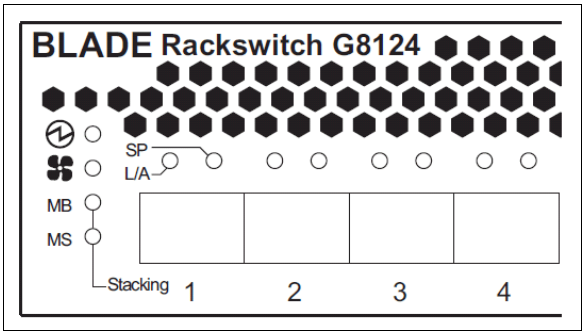

Figure 1-9 Location of LEDs on the RackSwitch G8124

Their meanings are explained in Figure 1-10.

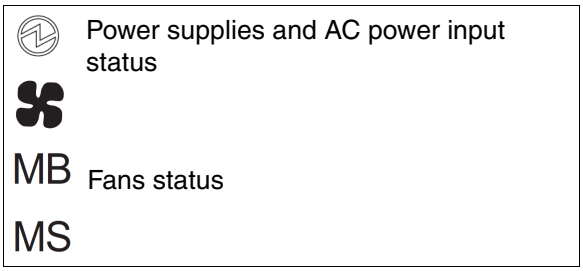

Figure 1-10 Indicator LEDs and their meanings

Table 1-6 shows the different System LED status for the RackSwitch G8124.

Table 1-6 RackSwitch G8124 System LED Status

Function	Power supply	Fans	Stacking master indicator[a]	Stacking member indicator[a]
Total Power Failure	Off	Off	Off	Off
Service Required	Flash green	Flash green	Flash green	Flash green
Power Supplies OK	Solid green	N/A	N/A	N/A
Power Supply Failure	Flash green	N/A	N/A	N/A
Fans OK	N/A	Solid green	N/A	N/A
Fan Failure	N/A	Flash green	N/A	N/A
Stack Master	N/A	N/A	On	Off
Stack Backup/Member	N/A	N/A	Off	On
Stack Error	N/A	N/A	On	N/A
Non-Stack Member	N/A	N/A	Off	Off

a. Stacking for the RackSwitch G8124 is not currently supported, but these indicators remain for possible future feature releases.

1.4.5 More Information

For more about the IBM System Networking RackSwitch G8124 and the LED status information, see the following resources:

- IBM System Networking RackSwitch G8124 Announcement Letter:

  http://www.ibm.com/common/ssi/cgi-bin/ssialias?infotype=AN&subtype=CA&appname=gpateam&supplier=899&letternum=ENUSLG11-0096&pdf=yes

- *IBM System Networking RackSwitch G8124*, TIPS0787
- *IBM System Networking RackSwitch G8124 Installation Guide*:

  https://www.ibm.com/support/docview.wss?uid=isg3T7000299&aid=1

- *IBM System Networking RackSwitch G8124/G8124E Application Guide*:

  http://www.ibm.com/support/docview.wss?uid=isg3T7000388

- *IBM System Networking RackSwitch G8124/G8124E Browser-Based Interface Quick Guide*:

  http://www-01.ibm.com/support/docview.wss?uid=isg3T7000389
- *IBM System Networking RackSwitch G8124/G8124E ISCLI Command Reference*:

  http://www-01.ibm.com/support/docview.wss?uid=isg3T7000390
- *IBM System Networking RackSwitch G8124/G8124E Menu-Based CLI Reference Guide*:

  http://www-01.ibm.com/support/docview.wss?uid=isg3T7000391

1.5 IBM System Networking RackSwitch G8264

The RackSwitch G8264 is a 10 Gb/40 Gb Top-of-Rack switch designed for applications that require the highest performance. It combines state of the art 1.28 Tbps throughput with up to 64 10 Gb SFP+ ports in an ultra-dense 1U form factor.

Figure 1-11 shows the front view of the RackSwitch G8264.

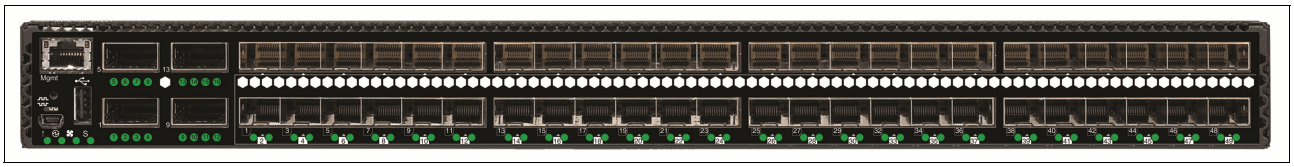

Figure 1-11 RackSwitch G8264 front view

RackSwitch G8264 is also designed to support IBM Virtual Fabric, which helps clients reduce cost and complexity when it comes to the I/O requirements of most virtualization deployments. Virtual Fabric can help clients reduce the number of multiple I/O adapters down to a single dual-port 10 G adapter, and reduce the number of cables and upstream switch ports required. Virtual Fabric allows clients to carve a dual-port 10 G server adapter into eight virtual NICs (vNICs) and create dedicated virtual pipes between the adapter and the switch for optimal performance, higher availability, and better security. With Virtual Fabric, IT can allocate bandwidth per vNIC and dynamically change, making room for the ability to adjust over time without downtime.

With the flexibility of the RackSwitch G8264 switch, clients can take advantage of the technologies that they require for multiple environments. The 40 Gb ports can also take advantage of the QSFP+ Break-out cables to gain an additional 16 x 10 Gb ports. For 40 Gb to 40 Gb connections, you can use QSFP+ direct-attached cables (DAC cables), which can be 1 - 3 meters in length and are ideal for connecting chassis together or connecting to another Top-of-Rack switch. For 40 Gb connectivity over longer distances, use a QSFP+ transceiver and QSFP+ MTP optical cables to connect up to 4 x 10 Gb ports using fiber (10 Gb transceivers are needed at the other end).

Table 1-7 shows the part numbers used to order the IBM System Networking RackSwitch G8264.

Table 1-7 IBM System Networking RackSwitch G8264 part numbers

Description	IBM Part No.	Power Systems MTM/FC
IBM System Networking RackSwitch G8264R (Rear-to-Front)	7309G64	1455-64C
IBM System Networking RackSwitch G8264F (Front-to-Rear)	730964F	

1.5.1 Switch inclusions

The module part numbers include the following items:

- One IBM System Networking RackSwitch G8264R/ G8264F
- Generic Rail Mount Kit (2-post)
- Mini-USB Console port for serial access

Transceivers: Small form-factor pluggable plus (SFP+) transceivers are not included in the purchase of the switch. All 1/10 Gb transceivers require LC-to-LC cables.

1.5.2 IBM System Networking RackSwitch G8264 benefits

The RackSwitch G8264 offers the following benefits:

- High performance: The 10 Gb/40 Gb Low Latency Switch provides the best combination of low latency, non-blocking line-rate switching, and ease of management. It also has a throughput of 1.2 Tbps.
- Lower power and better cooling: RackSwitch G8264 uses as little as 275 W of power, which is a fraction of the power consumption of most competitive offerings. Unlike side-cooled switches, which can cause heat recirculation and reliability concerns, the G8264 front-to-rear or rear-to-front cooling design reduces data center air conditioning costs by having airflow match the servers in the rack. In addition, variable speed fans assist in automatically reducing power consumption.
- Virtual Fabric: Virtual Fabric can help customers address I/O requirements for multiple NICs, while also helping reduce cost and complexity. Virtual Fabric for IBM allows for the carving up of a physical NIC into multiple virtual NICs (2 - 8 vNICs) and creates a virtual pipe between the adapter and the switch for improved performance, availability, and security, while reducing cost and complexity.
- VM-aware networking: VMready software on the switch simplifies configuration and improves security in virtualized environments. VMready automatically detects virtual machine movement between physical servers and instantly reconfigures each VM's network policies across VLANs to keep the network running without interrupting traffic or impacting performance. VMready works with all leading VM providers, such as VMware, Citrix, Xen, and Microsoft.
- Layer 3 functionality: The IBM System Networking RackSwitch includes Layer 3 functionality, which provides security and performance benefits, as inter-VLAN traffic stays within the switch. This switch also provides the full range of Layer 3 protocols from static routes for technologies, such as Open Shortest Path First (OSPF) and Border Gateway Protocol (BGP) for enterprise customers.

- Seamless interoperability: IBM System Networking RackSwitches interoperate seamlessly with other vendors' upstream switches.
- Fault tolerance: IBM System Networking RackSwitches learn alternative routes automatically and perform faster convergence in the unlikely case of a link, switch, or power failure. The switch uses proven technologies, such as L2 trunk failover, advanced VLAN-based failover, VRRP, and Hot Link.
- Multicast: Supports IGMP Snooping v1, v2, and v3 with 2 K IGMP groups, and Protocol Independent Multicast, such as PIM Sparse Mode or PIM Dense Mode.
- Converged fabric: The IBM System Networking RackSwitch is designed to support CEE and connectivity to FCoE gateways. CEE helps enable clients to combine storage, messaging traffic, VoIP, video, and other data on a common data center Ethernet infrastructure. FCoE helps enable highly efficient block storage over Ethernet for consolidating server network connectivity. As a result, clients can deploy a single server interface for multiple data types, which can simplify both deployment and management of server network connectivity, while maintaining the high availability and robustness required for storage transactions.

1.5.3 Features and specifications

In this section, we list some of the hardware and software features and specifications of the IBM System Networking RackSwitch G8264. For more details about these features, see Chapter 2, “IBM System Networking Switch 10Gb Ethernet switch features” on page 51.

Performance

The IBM System Networking RackSwitch G8264 has the following performance characteristics:

- 100% line rate performance
- 1280 Gbps non-blocking switching throughput (full duplex)
- Sub 1.1 microseconds latency
- 960 Mpps

Hardware features

The hardware features for the RackSwitch G8264 are:

- Models:
  - RackSwitch G8264R (for rear-to-front cooling). For ports, at the rear of the rack, matching System x and BladeCenter designs.
  - RackSwitch G8264F (for front-to-rear cooling). For ports, at the front of the rack, the matching airflow of iDataPlex.
- Interface options:
  - Forty-eight SFP+ ports (10 GbE).
  - Four QSFP+ ports (40 GbE).
  - One 10/100/1000 Ethernet RJ45 port for out-of-band management.
  - One USB port for storage device connection[3].

[3] If a USB drive is inserted into the USB port, you can copy files from the switch to the USB drive, or from the USB drive to the switch. You also can boot the switch by using software or configuration files found on the USB drive.

  - One mini-USB Console port for serial access, which provides an additional means to install software and configure the switch module. This USB-style connector enables connection of a special serial cable that is supplied with the switch module.
  - Server-like port orientations, enabling short and simple cabling.
- Dimensions: 17.3 in. wide, 19 in. deep, 1 RU high
- Weight: 9.98 kg (22 lb)
- Rack Installation Kit:
  - Generic Rack Mount Kit (2-post).
  - Optional versatile 4-post mounting options for 19-inch server rack or datacom rack.
  - Can be mounted vertically or horizontally.
- LEDs: System LEDs to indicate link status, fan, power, and stacking[4].
- Airflow:
  - Front-to-rear or rear-to-front cooling.
  - Redundant variable hot-swap speed fans for reduced power draw.
- Power:
  - Dual hot swap power modules, 50 - 60Hz, 100 - 240 VAC auto-switching per module.
  - Typical power consumption of 275 W.
- Environmental specifications:
  - Temperature: Ambient operating: 0 - 40 ºC.
  - Relative humidity: Non-condensing, operating 10 - 90%.
  - Altitude: Operating 3,050 m (10,000 feet).
  - Heat dissipation: 1127 BTU/hour (typical).
  - Mean time between failures (MTBF): 165,990 hours @ 40 °C.

Figure 1-12 shows the front view of the RackSwitch G8264 with the different ports that were described previously.

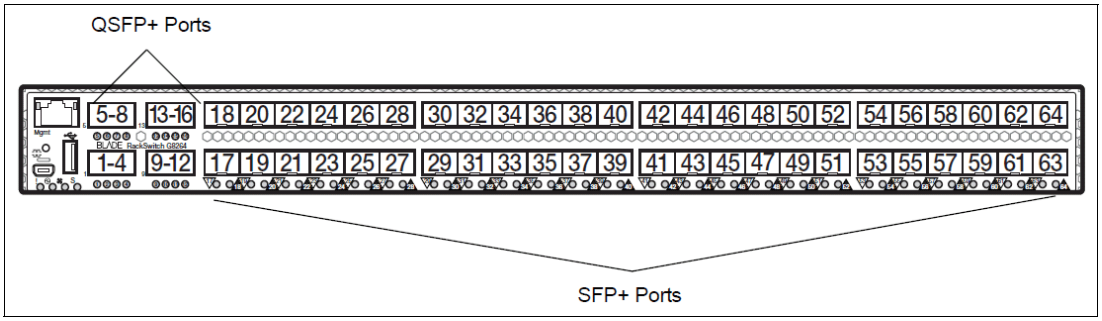

Figure 1-12 RackSwitch G8264 front panel

[4] Reserved for future OS release

Figure 1-13 shows the rear view of the switch.

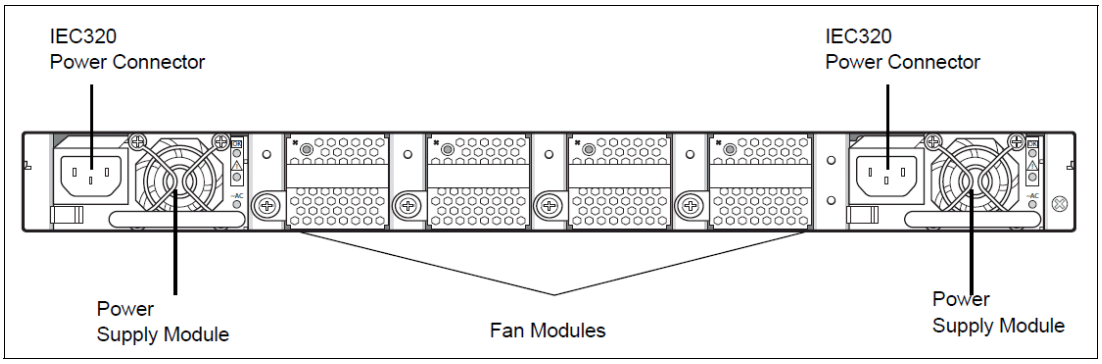

Figure 1-13 RackSwitch G8264 rear panel

Software features

The software features for the RackSwitch G8264 are:

- Security:
  - RADIUS
  - TACACS+
  - SCP
  - Wire Speed Filtering: Allow and Deny
  - SSH v1 and v2
  - HTTPS Secure BBI
  - Secure interface login and password
  - MAC address move notification
  - Shift B Boot menu (Password Recovery/ Factory Default)
- VLANs:
  - Port-based VLANs
  - 4096 VLAN IDs supported
  - 1024 Active VLANs (802.1Q)
  - 802.1x with Guest VLAN
  - Private VLAN Edge
- FCoE/Lossless Ethernet:
  - 802.1 Data Center Bridging
  - Priority Based Flow Control (PFC)
  - Enhanced Transmission Selection (ETS)
  - Data Center Bridge Exchange protocol (DCBX)
  - FIP Snooping
  - Fibre Channel over Ethernet (FCoE)
  - Converged Enhanced Ethernet (CEE)
- Trunking:
  - LACP
  - Static Trunks (EtherChannel)
  - Configurable Trunk Hash algorithm
- Spanning Tree:
  - Multiple Spanning Tree (802.1 s)
  - Rapid Spanning Tree (802.1 w)
  - PVRST+

  - Fast Uplink Convergence
  - BPDU guard
- Quality of Service:
  - QoS 802.1p (priority queues)
  - DSCP remarking
  - Metering
- Routing protocols:
  - RIP v1/v2
  - OSPF
  - BGP
- High availability:
  - Layer 2 failover
  - Hot Links
  - VRRP
- Multicast:
  - IGMP Snooping v1, v2, and v3 with 2 K IGMP groups
  - Protocol Independent Multicast (PIM Sparse Mode/Dense Mode)
- Monitoring:
  - Port mirroring
  - ACL-based mirroring
  - sFlow Version 5
- Virtualization:
  - VMready with VI API support
  - vNIC MIB support for SNMP
- Upgrades:
  - Upgrade firmware through serial or TFTP
  - Dual software images

Management features

RackSwitch G8264 supports the following management clients:

- IBM System Networking Element Manager
- isCLI (Cisco-like)
- Scriptable CLI
- Browser-based client or Telnet

Standard protocols

RackSwitch G8264 supports the following standard protocols:

- IPv6
- SNMP v1, v2c, and v3
- RMON
- Secondary NTP Support
- DHCP Client
- DHCP Relay
- LLDP
- 128 K MAC Table
- 9 K Jumbo Frames
- 802.3X Flow Control

1.5.4 IBM System Networking RackSwitch G8264 LED status details

Figure 1-14 shows the LED indicators as they appear on the switch.

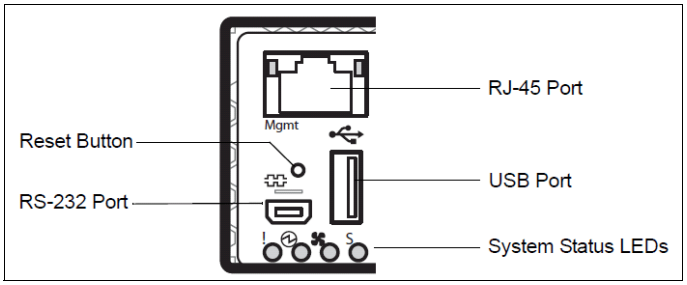

Figure 1-14 Location of LEDs on the RackSwitch G8264

Their meanings are explained in Figure 1-15.

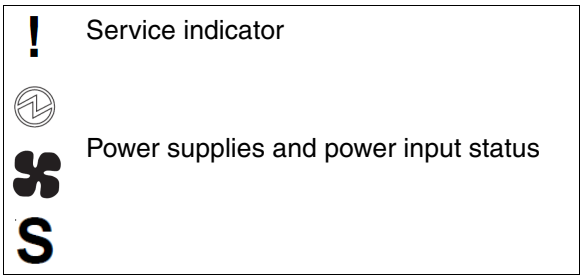

Figure 1-15 Indicator LEDs and their meanings

Table 1-8 shows the different System LED statuses for the RackSwitch G8264.

Table 1-8 RackSwitch G8264 System LED status

Function	Service	Power Supply	Fans	Stacking [a]
Total Power Failure	Off	Off	Off	Off
Service Required	Flash blue	Flash green[b]	Flash green[c]	Flash green or Solid green[d]
Power Supplies OK	N/A	Solid green	N/A	N/A
Power Supply Failure	Flash blue	Flash green	N/A	N/A
Fans OK	N/A	N/A	Solid Green	N/A
Fan Failure	Flash blue	N/A	Flash Green	N/A
Stack Master	Off	N/A	N/A	Flash green
Stack Backup/Member	On	N/A	N/A	Solid green
Stack Error	Flash blue	N/A	N/A	Flash green or Solid green
Non-Stack Member	Off	N/A	N/A	Off
Operations Command	Solid blue[e]	N/A	N/A	N/A

a. Stacking for the RackSwitch G8264 is not currently supported, but these indicators remain for possible future feature releases.

b. If service is required is because of a power supply failure, this LED Flash. Otherwise, it is solid green.
c. If service is required because of a fan failure, this LED Flash. Otherwise, it is solid green.
d. If service is required because of a stacking error, this LED Flash or is solid green, depending on its last known good state.
e. If an operations command is sent to the unit, this LED is solid blue. It can be used to locate the device.

1.5.5 More information

For more about the IBM System Networking RackSwitch G8264 and the LED status information, see the following resources:

- IBM System Networking RackSwitch G8264 Announcement Letter:

  http://www.ibm.com/common/ssi/cgi-bin/ssialias?infotype=AN&subtype=CA&appname=gpateam&supplier=872&letternum=ENUSAG11-0005&pdf=yes
- *IBM System Networking RackSwitch G8264/G8264T*, TIPS0815
- *IBM System Networking RackSwitch G8264 Installation Guide*:

  https://www-304.ibm.com/support/docview.wss?uid=isg3T7000294&aid=1
- *IBM System Networking RackSwitch G8264 Application Guide*:

  http://www.ibm.com/support/docview.wss?uid=isg3T7000326
- *IBM System Networking RackSwitch G8264 Browser-Based Interface Quick Guide*:

  http://www.ibm.com/support/docview.wss?uid=isg3T7000342
- *IBM System Networking RackSwitch G8264 ISCLI Command Reference*:

  http://www.ibm.com/support/docview.wss?uid=isg3T7000329
- *IBM System Networking RackSwitch G8264 Menu-Based CLI Reference Guide*:

  http://www-01.ibm.com/support/docview.wss?uid=isg3T7000328

1.6 IBM BladeCenter switches

This section describes the IBM BladeCenter switches, and then describes the two different chassis that these switches can be implemented in.

Ensure that you check the IBM ServerProven® website for compatibility information for blade servers, I/O adapters, and switches before implementation:

http://www-03.ibm.com/systems/info/x86servers/serverproven/compat/us/

1.6.1 IBM Virtual Fabric 10Gb Switch Module for IBM BladeCenter

The IBM Virtual Fabric 10Gb Switch Module for IBM BladeCenter offers the most bandwidth of any blade switch. It represents the perfect migration platform for clients who are still at 1 Gb either inside or outside the chassis by seamlessly integrating into the existing 1 Gb infrastructure. This switch is the first 10 Gb switch for IBM BladeCenter that supports converged networking (that is, the ability to transmit Converged Enhanced Ethernet (CEE) to a Fibre Channel over Ethernet (FCoE)-capable, Top-of-Rack switch). This feature is available with firmware release 6.1 and higher.

Using the CEE and FCoE functionality, you can transfer storage, network, Voice over IP (VoIP), video, and other data over the common Ethernet infrastructure. With the use of the QLogic Virtual Fabric Extension Module, clients can achieve FCoE gateway functionality inside the BladeCenter chassis.

The IBM Virtual Fabric 10Gb Switch Module can be used both in IBM Virtual Fabric Mode and Switch Independent Mode. The switch module can be managed by using a command-line interface (CLI) or web browser interface.

If you have a chassis with multiple servers, several operating at 1 Gbps, several at 10 Gbps, and several transmitting converged packets, this single switch can handle all of these workloads and can connect to a 1 Gb infrastructure, to a 10 Gb infrastructure, or both. Also after initial installation, this switch, along with all other IBM System Networking RackSwitches or IBM BladeCenter Switches, can be managed through IBM System Networking Element Manager.

With the extreme flexibility of this switch, you can take advantage of the technologies that are required for multiple environments. For 1 Gbps uplinks, they can take advantage of SFP transceivers. For 10 Gbps uplinks, they have a choice of either SFP+ transceivers (short range or long range for longer distances, or direct-attached copper (DAC) cables (also known as Twinax active cables) for shorter distances. DAC cables are more cost-effective, consume less power, and can be up to 7 m in length. They are ideal for connecting chassis together, connecting to a Top-of-Rack switch, or even connecting to an adjacent rack.

Figure 1-16 shows the IBM Virtual Fabric 10Gb Switch Module.

Figure 1-16 IBM Virtual Fabric 10Gb Switch Module

Table 1-9 lists the part number and feature code to use to order the module.

Table 1-9 IBM Virtual Fabric 10Gb Switch Module part number and feature code

Description	Part number	Feature code
IBM Virtual Fabric 10Gb Switch Module for IBM BladeCenter	46C7191	1639

The part number includes the following items:

- One IBM Virtual Fabric 10Gb Switch Module
- One 3 m mini-USB-to-DB9 serial console cable
- One filler module
- *Virtual Fabric 10Gb Switch Module Installation Guide*
- User license agreement

- Important Notices document
- Documentation CD-ROM

Transceivers: Small form-factor pluggable plus (SFP+) transceivers are not included in the purchase of the switch. All 1/10 Gb transceivers require LC-to-LC cables.

To communicate outside of the chassis, you must have either SFP+ transceivers or SFP+ direct-attach copper (DAC) cables connected. DAC cables have SFP+ transceivers on both ends. You have the flexibility to expand your bandwidth, using anywhere from one connection up to 10 connections per switch.

IBM Virtual Fabric 10Gb Switch Module features and functions

The IBM Virtual Fabric 10Gb Switch Module includes the following features and functions:

- Form-factor: Single-height, high-speed switch module
- Internal ports:
  - Fourteen internal auto-negotiating ports: 1 Gbps or 10 Gbps to the server blades.
  - Two internal, full-duplex 100Mbps ports connected to the management module.
- External ports:
  - Up to ten 10 Gb SFP+ ports (also designed to support 1 Gb SFP if required, with the flexibility to mix 1 Gb/10 Gb). The oversubscription ratio (14 internal ports to 10 external ports) is low, which makes the switch module suitable for the most performance-intensive environments
  - One 10/100/1000 Mb copper RJ45 that is used for management or data.
  - An RS-232 mini-USB connector for serial port that provides an additional means to install software and configure the switch module. This USB-style connector enables connection of a special serial cable that is supplied with the switch module.
- Scalability and performance:
  - Autosensing 1 Gb/10 Gb internal and external Ethernet ports for bandwidth optimization.
  - Non-blocking architecture with wire-speed forwarding of traffic and full line rate performance of 480 Gbps full duplex.
  - Media access control (MAC) address learning: Automatic updates, and supports up to 32 Kb MAC addresses.
  - Up to 128 IP interfaces per switch.
  - Static, EtherChannel, and Link Aggregation Control Protocol (LACP) (IEEE 802.3ad) link aggregation, up to 100 Gb of total bandwidth per switch, up to 18 trunk groups, and up to eight ports per group.
  - Support for jumbo frames (up to 12288 bytes).
  - Broadcast and multicast storm control.
  - IGMP snooping for limit flooding of IP multicast traffic (IGMP V1, V2, and V3).
  - Configurable traffic distribution schemes over trunk links, based on source and destination IP addresses, MAC addresses, or both.
  - Fast port forwarding and fast uplink convergence for rapid Spanning Tree Protocol (STP) convergence.
  - Stacking support: Clients can stack up to eight IBM Virtual Fabric 10Gb Switch Module.

- Availability and redundancy:
  - VRRP for Layer 3 router redundancy.
  - IEEE 802.1D STP for providing Layer 2 redundancy with PVRST+.
  - IEEE 802.1s Multiple STP (MSTP) for topology optimization, up to 128 STP instances supported by single switch.
  - IEEE 802.1w Rapid STP (RSTP), providing rapid STP convergence for critical delay-sensitive, traffic-like voice or video.
  - Layer 2 trunk failover to support active and standby configurations of network adapter teaming on blades.
  - Interchassis redundancy (Layer 2 and Layer 3).
- VLAN support:
  - Up to 1024 VLANs supported per switch, VLAN numbers that range 1 - 4095 (4095 is a dedicated VLAN used for the management module connection only).
  - 802.1Q VLAN tagging support on all ports.
  - Protocol-based VLANs.
- Security:
  - VLAN-based, MAC-based, and IP-based access control lists (ACLs).
  - 802.1X port-based authentication.
  - Multiple user IDs and passwords.
  - User access control.
  - Radius, Terminal Access Controller Access-Control System Plus (TACACS+), and Lightweight Directory Access Protocol (LDAP).
- Quality of Service (QoS):
  - Up to eight queues per port.
  - Support for IEEE 802.1p, IP ToS/DSCP, and ACL-based (MAC/IP source and destination addresses, VLANs) traffic classification and processing.
  - Traffic shaping and remarking based on defined policies.
  - Eight Weighted Round Robin (WRR) priority queues per port for processing qualified traffic.
- Layer 3 functions:
  - IP forwarding.
  - IP filtering with ACLs (up to 4096 ACLs supported).
  - VRRP for router redundancy.
  - Support for up to 128 static routes.
  - Routing protocol support (Router Information Protocol (RIP) V1, RIP V2, Open Shortest Path First protocol (OSPF)V1, V2, and V3, BGP-4), up to 1024 entries in routing table.
  - IPv6 routing, including static routes and OSPFv3 (requires firmware V6.3 or higher).
  - Support for Dynamic Host Configuration Protocol (DHCP) Relay.
  - IPv6 host management.
  - IPv6 forwarding based on static routes.

- Manageability:
  - Simple Network Management Protocol (SNMP V1, V2, and V3).
  - HTTP and HTTPS Browser-Based Interface (BBI).
  - Industry standard CLI and IBM Networking OS/AlteonOS CLI.
  - Telnet interface for CLI.
  - SSH v1/v2.
  - Serial interface for CLI.
  - Scriptable CLI.
  - Firmware image update (Trivial File Transfer Protocol (TFTP) and File Transfer Protocol (FTP)).
  - Network Time Protocol (NTP) for switch clock synchronization.
  - IBM System Networking Element Manager support.
- Monitoring:
  - Switch LEDs for external port status and switch module status indication.
  - Port mirroring for analyzing network traffic that passes through the switch.
  - Change tracking and remote logging with syslog feature.
  - Power-On Self Test (POST) diagnostic tests.
- Special functions: Serial over LAN (SOL).
- Virtualization features:
  - VMready.
  - Virtual Fabric Adapter vNIC support.
- Converged Enhanced Ethernet and FCoE features:
  - FCoE allows Fibre Channel traffic to be transported over Ethernet links.
  - FCoE Initialization Protocol (FIP) snooping to enforce point-to-point links for FCoE traffic outside the regular Fibre Channel topology.
  - Priority-Based Flow Control (PFC) (IEEE 802.1Qbb) extends the 802.3x standard flow control to allow the switch to pause traffic, based on the 802.1p priority value in each packet VLAN tag.
  - Enhanced Transmission Selection (ETS) (IEEE 802.1Qaz) provides a method for allocating link bandwidth, based on the 802.1p priority value in each packet VLAN tag.
  - DCBX (IEEE 802.1AB) allows neighboring network devices to exchange information about their capabilities.
  - Supports the QLogic Virtual Fabric Extension Module for IBM BladeCenter, which provides FCoE gateway functionality inside the BladeCenter chassis.

VMready is a unique solution that enables the network to be virtual machine aware. The network can be configured and managed for virtual ports (v-ports), rather than just for physical ports. With VMready, as VMs migrate across physical hosts, so do their network attributes. Virtual machines can be added, moved, and removed, while retaining the same ACLs, QoS, and VLAN attributes. VMready allows for a *define-once-use-many* configuration that evolves as the server and network topologies evolve. VMready works with all virtualization products, including VMware, Hyper-V, Xen, and KVM, without modification of virtual machine hypervisors or guest operating systems. It is available as part of release 6.1 (and higher).

VMready compatibility with Virtual Fabric solutions is as follows:

- VMready is not supported with IBM Virtual Fabric Mode.
- VMready is supported with Switch Independent Mode.

The switch module supports the following IEEE standards:

- IEEE 802.1D STP with PVRST+
- IEEE 802.1s MSTP
- IEEE 802.1w RSTP
- IEEE 802.1p Tagged Packets
- IEEE 802.1Q Tagged VLAN (frame tagging on all ports when VLANs are enabled)
- IEEE 802.1x port-based authentication
- IEEE 802.2 Logical Link Control
- IEEE 802.3ad Link Aggregation Control Protocol
- IEEE 802.3x Full-duplex Flow Control
- IEEE 802.3ab 1000BASE-T Gigabit Ethernet
- IEEE 802.3ae 10GBASE-SR 10Gb Ethernet fiber optics short range
- IEEE 802.3ae 10GBASE-LR 10Gb Ethernet fiber optics long range
- IEEE 802.3z 1000BASE-SX Gigabit Ethernet

The following network cables are supported for the IBM Virtual Fabric 10Gb Switch Module:

- 10GBASE-SR for 10Gb ports: 850 nm wavelength, multimode fiber, 50µ or 62.5µ (300 m maximum), with LC duplex connector
- 1000BASE-T for RJ45 port:
  - UTP Category 6 (100 m maximum)
  - UTP Category 5e (100 m maximum)
  - UTP Category 5 (100 m maximum)
  - EIA/TIA-568B 100-ohm STP (100 meters maximum)

Ensure that you check the IBM ServerProven site for compatibility between blade servers, I/O adapters, and switches before implementation:

http://www.ibm.com/systems/info/x86servers/serverproven/compat/us/

More information

For more information about the IBM Virtual Fabric 10Gb Switch Module for IBM BladeCenter, see the following resources:

- IBM Virtual Fabric 10Gb Switch Module Announcement Letter:

  http://www.ibm.com/common/ssi/rep_ca/5/872/ENUSAG09-0245/ENUSAG09-0245.PDF

- *BNT Virtual Fabric 10Gb Switch Module for IBM BladeCenter*, TIPS0708
- *IBM Virtual Fabric 10Gb Switch Module for IBM BladeCenter Installation Guide*:

  http://download.boulder.ibm.com/ibmdl/pub/systems/support/system_x_pdf/46m1525.pdf

- *IBM Virtual Fabric 10Gb Switch Module for IBM BladeCenter Application Guide*:

  http://download.boulder.ibm.com/ibmdl/pub/systems/support/system_x_pdf/bmd00189.pdf

- *IBM Virtual Fabric 10Gb Switch Module for IBM BladeCenter Command Reference*:

  http://download.boulder.ibm.com/ibmdl/pub/systems/support/system_x_pdf/bmd00190.pdf

- *IBM Virtual Fabric 10Gb Switch Module for IBM BladeCenter isCLI Reference*:

  http://download.boulder.ibm.com/ibmdl/pub/systems/support/system_x_pdf/bmd00191.pdf

- *IBM Virtual Fabric 10Gb Switch Module for IBM BladeCenter BBI Quick Guide*:

  http://download.boulder.ibm.com/ibmdl/pub/systems/support/system_x_pdf/bmd00192.pdf

1.6.2 IBM 1/10 Uplink Ethernet Switch Module for IBM BladeCenter

The IBM 1/10 Uplink Ethernet Switch Module for IBM BladeCenter is a switch option that enables administrators to offer full Layer 2 and 3 switching and routing capability with combined 1 Gb and 10 Gb uplinks in a BladeCenter chassis. Such consolidation simplifies the data center infrastructure and helps reduce the number of discrete devices, management consoles, and management systems. In addition, the next-generation switch module hardware can support IPv6 Layer 3 frame forwarding protocols by using a future firmware upgrade.

This Ethernet switch module delivers port flexibility, efficient traffic management, increased uplink bandwidth, and strong Ethernet switching price/performance. Figure 1-17 shows the switch module.

Figure 1-17 IBM 1/10 Uplink Ethernet Switch Module view

Table 1-10 lists the part number and feature code to use to order the module.

Table 1-10 IBM 1/10 Uplink Ethernet Switch Module part number and feature code

Description	Part number	Feature code
IBM 1/10 Uplink Ethernet Switch Module for IBM BladeCenter	44W4407	1590 / 6980 / 1590[a]

a. Feature codes are listed in the form of three codes separated by a forward slash mark (/). The first feature code is for BladeCenter E-, T-, H-, and HT based configurations that are available through the IBM System x platform. The second feature code is for BladeCenter S-based configurations that are available through the System x sales channel. The third feature code is for BladeCenter S- and BladeCenter H-based configurations that are available through the IBM Power Systems sales channel when applicable.

The part number includes the following items:

- One IBM 1/10 Uplink Ethernet Switch Module for IBM BladeCenter
- 3 m USB-to-DB9 serial console cable
- Printed documentation
- Documentation CD-ROM

Features and specifications

The IBM 1/10 Uplink Ethernet Switch Module includes the following standard features and functions:

- Internal ports:
  - 14 internal full-duplex Gigabit ports, one connected to each of the blade servers.
  - Two internal full-duplex 10/100 Mbps ports connected to the management module.
  - External ports:
    - Three slots for 10 Gb Ethernet SFP+ transceivers (support for 10GBASE-SR or 10GBASE-LR) or SFP+ copper direct-attach cables (DAC); SFP+ modules are optional
    - Six 10/100/1000 1000BASE-T Gigabit Ethernet ports with RJ-45 connectors
    - One RS-232 serial port that provides an additional means to install software and configure the switch module
- Scalability and performance:
  - Fixed-speed external 10 Gb Ethernet ports for maximum uplink bandwidth.
  - Autosensing 10/1000/1000 external Gigabit Ethernet ports for bandwidth optimization.
  - Non-blocking architecture with wire-speed forwarding of traffic.
  - Media access control (MAC) address learning: Automatic update. Supports up to 16 K MAC addresses.
  - Up to 128 IP interfaces per switch.
  - Static and LACP (IEEE 802.3ad) link aggregation, up to 36 Gb of total bandwidth per switch, up to 16 trunk groups, up to six ports per group.
  - Support for jumbo frames (up to 9216 bytes).
  - Broadcast/multicast storm control.
  - IGMP snooping for limit flooding of IP multicast traffic.
  - IGMP filtering to control multicast traffic for hosts that participate in multicast groups.
  - Configurable traffic distribution schemes over trunk links based on source/destination IP or MAC addresses or both.
  - Fast port forwarding and fast uplink convergence for rapid STP convergence.
- Availability and redundancy:
  - VRRP for Layer 3 router redundancy.
  - IEEE 802.1D STP for providing L2 redundancy.
  - IEEE 802.1s Multiple STP (MSTP) for topology optimization. Up to 128 STP instances are supported by single switch.
  - IEEE 802.1w Rapid STP (RSTP) provides rapid STP convergence for critical delay-sensitive traffic like voice or video.

  - Layer 2 Trunk Failover to support active/standby configurations of network adapter teaming on blades.
  - Interchassis redundancy (L2 and L3).
- VLAN support:
  - Up to 1024 VLANs supported per switch, with VLAN numbers ranging 1 - 4095 (4095 is used for management module's connection only).
  - 802.1Q VLAN tagging support on all ports.
  - Private VLANs.
- Security:
  - VLAN-based, MAC-based, and IP-based ACLs.
  - 802.1X port-based authentication.
  - Multiple user IDs and passwords.
  - User access control.
  - Radius/TACACS+.
- Quality of Service (QoS):
  - Support for IEEE 802.1p, IP ToS/DSCP, and ACL-based (MAC/IP source and destination addresses, VLANs) traffic classification and processing.
  - Traffic shaping and remarking based on defined policies.
  - Eight Weighted Round Robin (WRR) priority queues per port for processing qualified traffic.
- Layer 3 functions:
  - IP forwarding.
  - IP filtering with ACLs. Up to 896 ACLs supported.
  - VRRP for router redundancy.
  - Support for up to 128 static routes.
  - Routing protocol support (RIP v1, RIP v2, OSPF v2, BGP-4). Up to 2048 entries in a routing table.
  - Support for DHCP Relay.
  - IPv6 host management support.
- Manageability:
  - Simple Network Management Protocol (SNMP V1 and V3).
  - HTTP browser GUI.
  - Telnet interface for CLI.
  - SSH.
  - Serial interface for CLI.
  - Scriptable CLI.
  - Firmware image update (TFTP and FTP).
  - Network Time Protocol (NTP) for switch clock synchronization.
- Monitoring:
  - Switch LEDs for external port status and switch module status indication.
  - Port mirroring for analyzing network traffic passing through switch.

- Change tracking and remote logging with syslog feature.
- POST diagnostic tests.

- Special functions: Support for Serial over LAN (SOL)
- Standards supported:

  The switch module supports the following IEEE standards:
  - IEEE 802.1D Spanning Tree Protocol (STP).
  - IEEE 802.1s Multiple STP (MSTP).
  - IEEE 802.1w Rapid STP (RSTP).
  - IEEE 802.1p Tagged Packets.
  - IEEE 802.1Q Tagged VLAN (frame tagging on all ports when VLANs are enabled).
  - IEEE 802.1x port-based authentication.
  - IEEE 802.2 Logical Link Control.
  - IEEE 802.3 10BASE-T Ethernet.
  - IEEE 802.3u 100BASE-TX Fast Ethernet.
  - IEEE 802.3ab 1000BASE-T Gigabit Ethernet.
  - IEEE 802.3z 1000BASE-X Gigabit Ethernet.
  - IEEE 802.3ad Link Aggregation Control Protocol.
  - IEEE 802.3x Full-duplex Flow Control.
  - IEEE 802.3ae 10GBASE-SR.

Connectors and LEDs

Figure 1-18 shows the front panel of the IBM 1/10 Uplink Ethernet Switch Module.

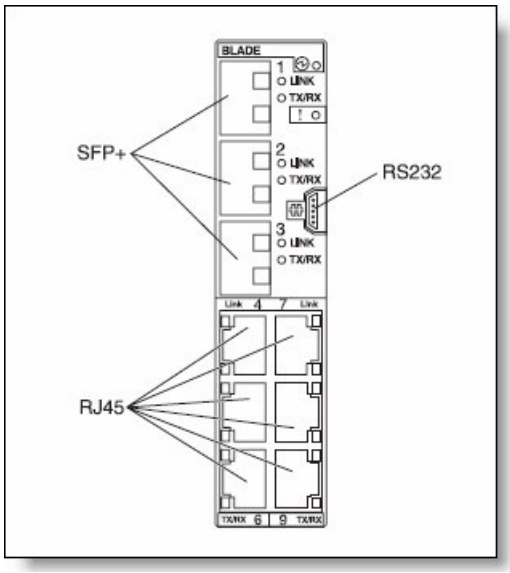

Figure 1-18 IBM 1/10 Uplink Ethernet Switch Module Front panel that shows connector and LED locations

The front panel contains the following components:

- LEDs that display the status of the switch module and the network. These LEDs are OK, which indicates that the switch module passed the power-on self-test (POST) with no critical faults and is operational, and switch module error, which indicates that the switch module failed the POST or detected an operational fault.
- One USB RS-232 console port that provides an additional means to install software and configure the switch module. This USB-style connector enables connection of a special serial cable that is supplied with the switch module.
- Six external 1000BASE-T Ethernet ports for 10/100/1000 Mbps connections to external Ethernet devices.
- Three external SFP+ port connectors to attach SFP+ modules for 1000 Mbps connections to external Ethernet devices.
- An Ethernet link OK LED and an Ethernet Tx/Rx LED for each external port on the switch module.

Network cabling requirements

The network cables required for the switch module are as follows:

- 10GBASE-SR:
  - 850 nm communication using multimode fiber cable (50µ or 62.5µ) up to 300 m
  - Requires 10GbE SFP+ transceiver modules, part number 44W4408, feature code 4942 (no SFP+ modules are shipped standard with the switch module)
- 10BASE-T:
  - UTP Category 3, 4, 5 (100 m (328 feet) maximum)
  - 100-ohm STP (100 m maximum)
- 100BASE-TX:
  - UTP Category 5 (100 m maximum)
  - EIA/TIA-568 100-ohm STP (100 m maximum)
- 1000BASE-T:
  - UTP Category 6
  - UTP Category 5e (100 m maximum)
  - UTP Category 5 (100 m maximum)
  - EIA/TIA-568B 100-ohm STP (100 m maximum)
- RS-232 serial cable: 3 m console cable DB-9-to-USB connector (nonstandard use of USB connector) that comes with the GbE switch module.

Ensure that you check the IBM ServerProven site for compatibility between blade servers, I/O adapters, and switches before implementation:

http://www-03.ibm.com/systems/info/x86servers/serverproven/compat/us/

More information

For more information about IBM BladeCenter, see the following resources:

- IBM 1/10Gb Uplink Ethernet Switch Module Announcement Letter:

  http://www.ibm.com/common/ssi/rep_ca/5/872/ENUSAG08-0365/ENUSAG080365.PDF
- *BNT 1/10Gb Uplink Ethernet Switch Module for IBM BladeCenter*, TIPS0705

- *IBM 1/10Gb Uplink Ethernet Switch Module for IBM BladeCenter Installation Guide*:

  ftp://ftp.software.ibm.com/systems/support/system_x_pdf/dw1gymst.pdf

- *IBM 1/10Gb Uplink Ethernet Switch Module for IBM BladeCenter Application Guide*:

  http://www-947.ibm.com/systems/support/supportsite.wss/docdisplay?brandind=5000008&lndocid=MIGR-5076214

- *IBM 1/10Gb Uplink Ethernet Switch Module for IBM BladeCenter Command Reference*:

  http://www-947.ibm.com/systems/support/supportsite.wss/docdisplay?brandind=5000008&lndocid=MIGR-5076525

- *IBM 1/10Gb Uplink Ethernet Switch Module for IBM BladeCenter isCLI Reference*:

  http://www-947.ibm.com/systems/support/supportsite.wss/docdisplay?brandind=5000008&lndocid=MIGR-5076215

- *IBM 1/10Gb Uplink Ethernet Switch Module for IBM BladeCenter BBI Quick Guide*:

  http://www-947.ibm.com/systems/support/supportsite.wss/docdisplay?brandind=5000008&lndocid=MIGR-5076219

1.6.3 IBM BladeCenter H

The IBM BladeCenter H (Figure 1-19) is a high-performance 9U chassis, designed for compute-intensive environments, such as Earth/Life Sciences, commercial analytics, and next-generation network applications.

Figure 1-19 IBM BladeCenter H Chassis front view

It provides:

- Reduced single points of failure: Many major components (either standard or optionally) are hot-swappable or redundant. Servers and modules can be configured for automatic failover to backups.
- Compatibility with earlier versions: Every blade, switch, and pass-through module released by IBM, since the original IBM BladeCenter E chassis in 2002, is supported in the IBM BladeCenter H chassis.

- High-speed redundant midplane connections: Based on 4X InfiniBand, the midplane supports up to 40 Gb bandwidth and provides four 10 Gb data channels to each blade. By giving each blade two physical connections to the midplane that connects all blades and modules together internally, a failure of one connector alone cannot bring down the server.
- Fourteen 30 mm blade slots: These hot-swap slots can support any combination of 14 blade servers, or seven double-wide (60 mm) blade servers, or a mixture of 30 mm and 60 mm blades. It also supports multiple optional 30 mm Expansion Units in combination with the blade servers, using the same blade slots. Up to four chassis can be installed in an industry-standard 42U rack, for a total of up to 56 30 mm blade servers per rack.
- Up to 10 module bays for communication and I/O switches or bridges: The modules interface with all of the blade servers in the chassis and alleviate the need for external switches or expensive, cumbersome, and error-prone cabling. All connections are done internally through the midplane. Two module slots are reserved for hot-swap/redundant Gigabit Ethernet switch modules. Two slots support either high-speed bridge modules or traditional Gigabit Ethernet, Myrinet, Fibre Channel, InfiniBand, and other switch modules. Two slots are dedicated for bridge modules. Four additional slots are dedicated for hot-swap/redundant high-speed switch modules. All modules, when installed in pairs, offer load balancing and failover support.

  Figure 1-20 shows the module bay locations.

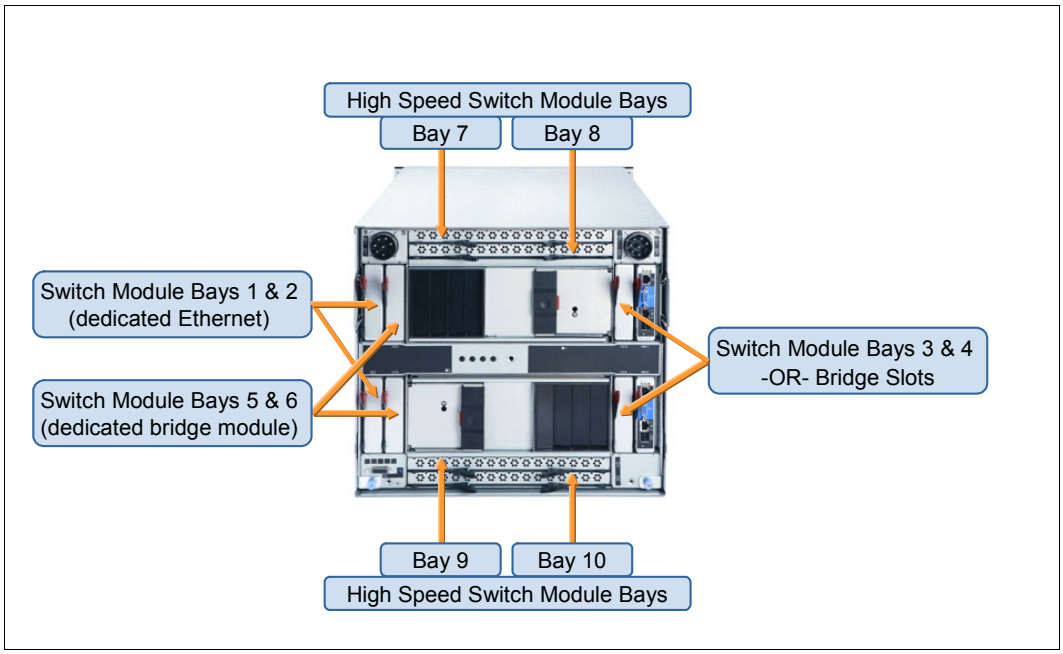

Figure 1-20 IBM BladeCenter H chassis switch module identification

- Integrated switch and bridge modules: No additional rack U space is required.
- Two module bays for Advanced Management Modules (AMMs): The management modules provide advanced systems management and KVM capabilities for not only the chassis itself, but for all of the blades and other modules installed in the chassis. The AMM provides capabilities similar to the Baseboard Management Controller used in stand-alone System x rack and tower servers. Features include concurrent KVM (cKVM), an external Serial over LAN connection, industry-standard management interfaces (SMASH/CLP/CIM/HPI), USB virtualization, network failover, compatibility with an earlier version of the original Management Module, and so on.

The features of the module can be accessed either locally or remotely across a network. One module comes standard. A second module can be added for hot-swap/redundancy and failover. The module uses USB ports for keyboard and mouse.

- Two module bays for blower modules: Two hot-swap/redundant blower modules come standard with the chassis. They can provide efficient cooling for up to 14 blades. These modules replace the need for each blade and switch to contain its own fans. The blowers are more energy efficient than dozens or hundreds of smaller fans would be, and they offer fewer points of potential failure. IBM BladeCenter H also includes up to four additional hot-swap/redundant fan packs to cool the power supplies and high-speed switch modules.
- Four bays for Power Modules: IBM BladeCenter H ships with two 2980 W high-efficiency hot-swap/redundant power modules (upgradeable to four), capable of handling the power needs of the entire chassis, including future higher-wattage processors. Each power module includes a customer-replaceable hot-swap/redundant fan pack (three fans) for additional cooling capability.

Power modules: Two additional power modules are required when more than six blades *or* high-speed switches are installed.

- A hot-swappable Media Tray containing a DVD-ROM drive, two USB 2.0 ports, and a light path diagnostic panel: The media tray is shared by all the blades in the server. This setup reduces unnecessary parts (and reduces the number of parts that can fail). If there is a failure of the Media Tray, the tray can be swapped for another tray. While the tray is offline, the servers in the chassis can remotely access the Media Tray in another chassis. The light path diagnostic panel contains LEDs that indicate chassis status.

Ethernet switch modules, network interfaces, and port designation in IBM BladeCenter H

It is important to understand how Blade Server options and their associated ports are allocated within the BladeCenter Chassis and which ports, within a blade server, map to the different I/O bays within the BladeCenter chassis.

Blade server default ports: Ethernet

Figure 1-21 shows a blade server in an IBM BladeCenter H chassis. The example blade shows two onboard NICs (NIC0 and NIC1). NIC0 is connected through the IBM BladeCenter H midplane to switch bay one. NIC1 is connected to switch bay two. These connections are hardwired in the chassis mid-plane and cannot be changed. Switch bays one and two in the IBM BladeCenter H are dedicated Ethernet switch bays; only Ethernet capable devices can be installed in these two switch bays.

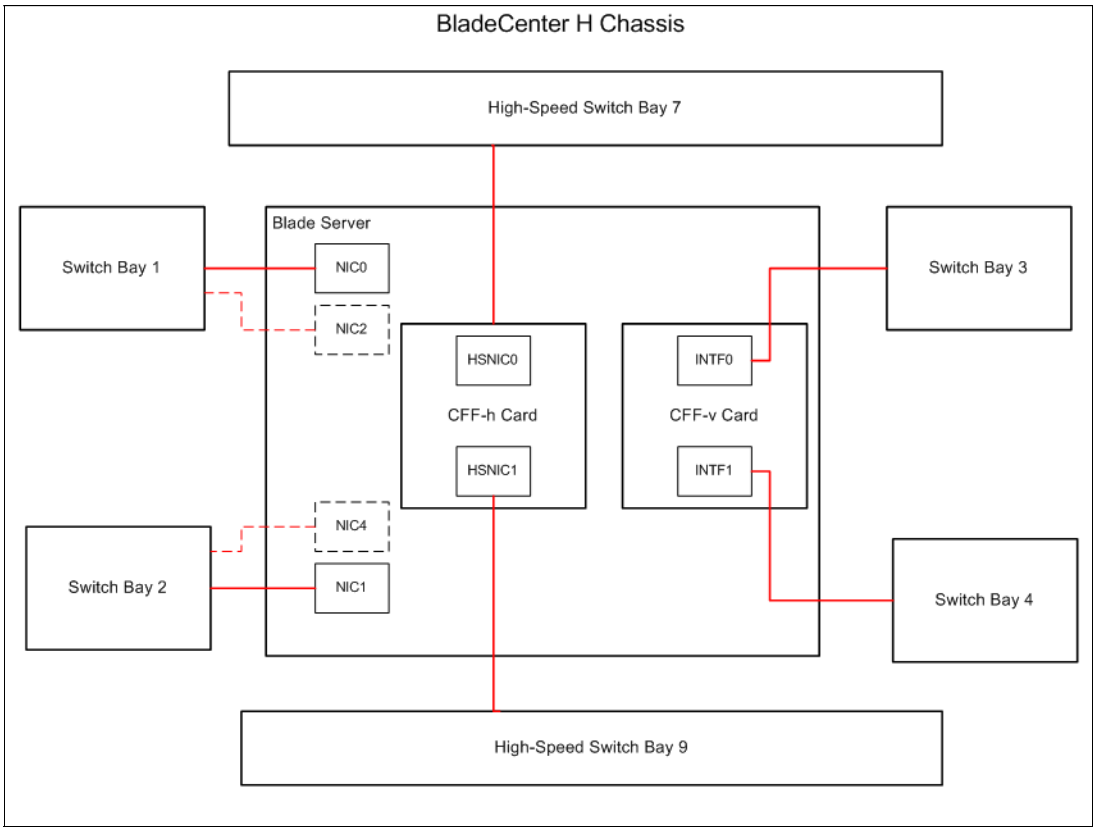

Figure 1-21 BladeCenter H Ethernet Switch Module and blade server NIC connections

Some blade servers, such as the 4 Socket Double-Wide HX5 (MT7873), are 60 mm wide and occupy two server slots in the chassis. The 4 Socket Double-Wide HX5 (and other double-wide blades) have an additional pair of Ethernet interfaces on the second module of the blade. These additional NICs (depicted as NIC2 and NIC4) are also hardwired to switch bays one and two.

Blade server additional ports: 1 Gb Ethernet / Fibre Channel

Switch bays three and four are optional switch bays. These switch bays can be used for Ethernet connectivity by installing an optional Combination I/O Vertical (CIO-v) Ethernet daughter card on the blade server. Interface 0 of the daughter card (INTF0) is connected to switch bay three and interface 1 (INTF1) is connected to switch bay four. Again, these connections cannot be changed. The appropriate Ethernet switch modules must also be installed in switch bays three and four to support Ethernet connectivity to the CIO-v Ethernet card.

Blade server high speed ports: 10 Gb Ethernet / 1 Gb Ethernet / Fibre Channel

Switch bays seven to ten are optional high-speed switch bays. These switch bays are used for additional Ethernet connectivity by installing the IBM Virtual Fabric 10Gb Switch Module into the appropriate high-speed switch bays and an optional Combo Form Factor Horizontal (CFF-h) High Speed Ethernet daughter card on the blade server. If a two (or four) port daughter card is installed, then Interface 0 of the daughter card (HSNIC0) is connected to switch bay seven, and interface 1 (HSNIC1) is connected to switch bay nine. If a four port card is installed in the Blade Server, then additional switches are required to use those additional ports. Again, these connections are hardwired from the server and cannot be changed.

Ensure that you check the IBM ServerProven site for compatibility between blade servers, I/O adapters, and switches before implementation:

http://www-03.ibm.com/systems/info/x86servers/serverproven/compat/us/

More information

For more information about, the IBM BladeCenter H, see the following resources:

- IBM BladeCenter H Announcement Letter:

  http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?infotype=an&subtype=ca&appname=gpateam&supplier=897&letternum=ENUS109-438

- *IBM BladeCenter Products and Technology*, SG24-7523
- *IBM BladeCenter H Installation and Users Guide*:

  http://publib.boulder.ibm.com/infocenter/bladectr/documentation/topic/com.ibm.bladecenter.8852.doc/bc_8852_iug.html

- *IBM BladeCenter H Trouble Shooting*:

  http://publib.boulder.ibm.com/infocenter/bladectr/documentation/topic/com.ibm.bladecenter.8852.doc/bc_8852_pdsg.html

1.6.4 IBM BladeCenter HT

IBM BladeCenter HT (Figure 1-22) is a carrier grade, rugged 12U chassis designed for challenging central office and networking environments.

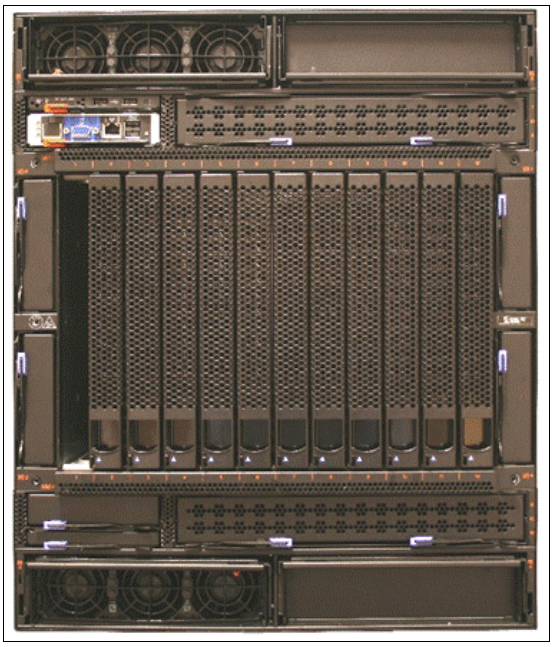

Figure 1-22 IBM BladeCenter HT chassis front view

It provides:

- NEBS Level 3/ETSI-tested: Designed for the Network Equipment Provider (NEP) and Service Provider (SP) environments. Also ideal for government, military, aerospace, industrial automation and robotics, medical imaging, and finance. Certified testing by Underwriters Laboratories of the IBM BladeCenter HT chassis is in progress; when complete, it will be covered under a UL-certified NEBS Level 3/ETSI test report.
- Designed for Carrier-Grade Linux: Several distributions are supported, include SUSE and Red Hat.
- Reduced single points of failure: Many major components (either standard or optionally) are hot-swappable or redundant. Servers and modules can be configured for automatic failover to backups.
- Compatibility with earlier versions: Every blade, switch, and pass-through module released by IBM, since the original IBM BladeCenter E chassis in 2002, is supported in the IBM BladeCenter HT chassis.
- High-speed redundant midplane connections: Based on 4X InfiniBand, the midplane supports up to 40 Gb bandwidth and provides four 10 Gb data channels to each blade. By giving each blade two physical connections to the midplane that connects all blades and modules together internally, a failure of one connector alone cannot bring down the server.
- Twelve 30 mm blade slots: These hot-swap slots can support any combination of blade servers, or six double-wide (60 mm) blade servers, or a mixture of 30 mm and 60 mm blades. It also supports multiple optional 30 mm Expansion Units in combination with the blade servers, using the same blade slots. Up to three chassis can be installed in an industry-standard 42U rack, for a total of up to 36 30 mm blade servers per rack.

- Two module bays for Advanced Management Modules: The management modules provide advanced systems management and KVM capabilities for not only the chassis itself, but for all of the blades and other modules installed in the chassis.
- Four bays for Fan Modules: All four hot-swap/redundant fan modules come standard with the chassis. These modules replace the need for each blade to contain its own fans. The high availability modules are more energy efficient than dozens or hundreds of smaller fans would be, and there are fewer points of potential failure.
- Four bays for Power Modules: IBM BladeCenter HT ships with two high-efficiency hot-swap/redundant DC or AC (model-specific) power modules (upgradeable to four), capable of handling the power needs of up to six blade servers.

Note: Two additional power modules are required when more than six blades *or* high-speed switches are installed.

- Two hot-swappable Media Trays: Each contain two external USB 2.0 ports, a light path diagnostic panel, and support a 1 Gb/4 Gb compact flash (CF) option: The media tray is shared by all the blades in the server. This setup reduces unnecessary parts (and reduces the number of parts that can fail). If there is a failure of the Media Tray, the tray can be swapped for another tray. While the tray is offline, the servers in the chassis can remotely access the Media Tray in another chassis. The light path diagnostic panel contains LEDs that indicate chassis status. One media tray comes standard (without compact flash); an optional second one provides redundancy. The CF option can act as a boot device, eliminating the need for HDDs in the blades.
- Redundant midplane connections: Each chassis contains a midplane that connects all blades and modules together internally. The midplane provides two physical connections to each blade; therefore, a failure of one connector alone cannot bring down the server.

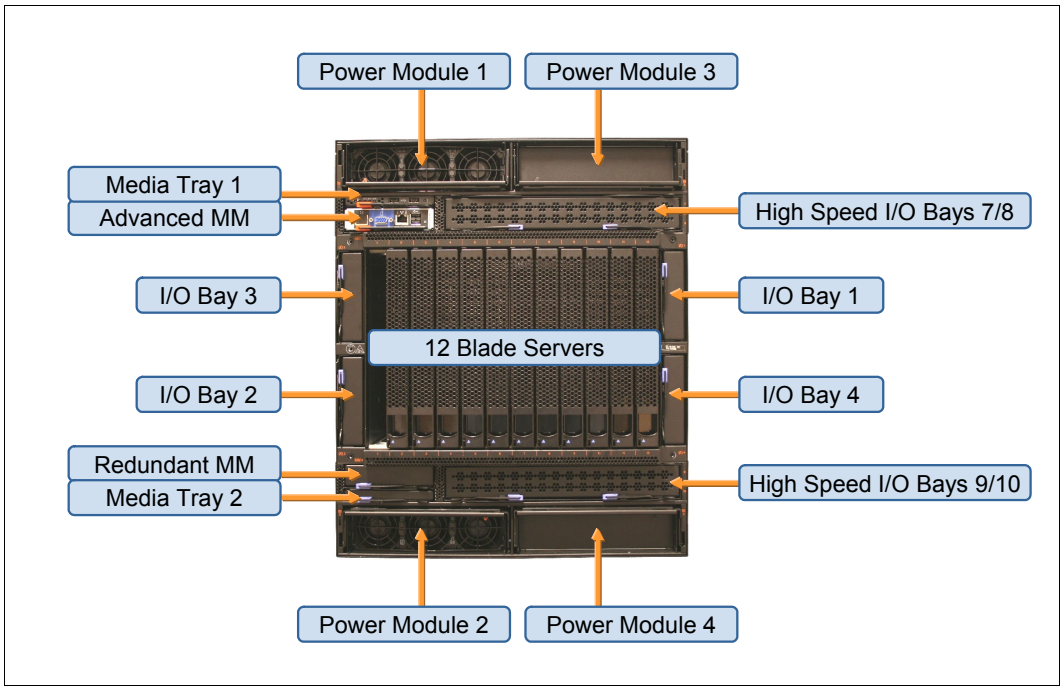

Figure 1-23 IBM BladeCenter HT chassis module locations

- Up to eight module bays for communication and I/O switches or bridges: The modules interface with all of the blade servers in the chassis and alleviate the need for external switches or expensive and cumbersome cabling. All connections are done internally through the midplane. Two module slots are reserved for hot-swap/redundant Gigabit Ethernet switch modules. Two slots support either high-speed bridge modules or traditional Gigabit Ethernet, Myrinet, Fibre Channel, InfiniBand, and other switch modules. Four additional slots are dedicated for hot-swap/redundant high-speed switch modules. All modules, when installed in pairs, offer load balancing and failover support.
- Bridge module: Dedicated bridge module bays do not exist in IBM BladeCenter HT. I/O bays 3 and 4 are similar to IBM BladeCenter H, but bridge bays 5 and 6 do not exist. IBM BladeCenter HT I/O bays 1 and 2 do provide bridge module support, but no hybrid bridge/Gb switch modules are planned at this time.
- Interposers: AMMs and I/O (switch) modules are all in the front of the chassis with the blades. Because they plug into a common midplane with the blade servers, they require interposers in each populated I/O bay to make up the difference in depth between switch enclosures and blade server enclosures. These interposers are entirely passive, and extend the midplane out to the rear connectors of the I/O modules. Figure 1-24 shows the interposer.

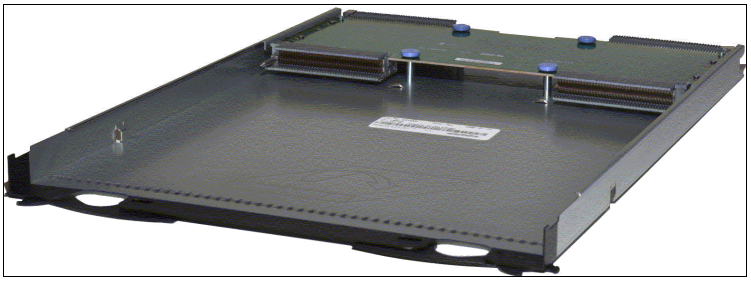

Figure 1-24 IBM BladeCenter HT Interposer - used for extending the midplane to connect the high speed I/O modules.

- **I**nterswitch Links: Because IBM BladeCenter HT has two unimplemented blade ports (13 and 14), the IBM BladeCenter HT midplane provides interconnects between switch ports 13 and 14 for the following matched pairs of switch modules: 1-2, 3-4, 7-9, 8-10. These pairs must be enabled by specifically designed switch interposers that complete the link between the switch module ports 13 and 14 and the midplane paths. If matching interposers do not both have the ISL support, then the switch module ports is not connected. Interswitch links between matched pairs of switch modules allow additional functionality, such as link aggregation or stacking, if supported.

Ethernet switch modules, network interfaces, and port designation in the IBM BladeCenter HT

It is important to understand how blade server options and their associated ports are allocated within the BladeCenter chassis and which ports, within a blade server, map to the different I/O bays within the BladeCenter chassis. There are additional switch connection considerations when using the IBM BladeCenter HT.

Blade server default ports: Ethernet

Figure 1-25 shows a blade server in an IBM BladeCenter HT chassis. The example blade shows two onboard NICs (NIC0 and NIC1). NIC0 is connected through the IBM BladeCenter HT midplane to switch bay one. NIC1 is connected to switch bay two. These connections are hardwired in the chassis mid-plane and cannot be changed. Switch bays one and two in the IBM BladeCenter HT are dedicated Ethernet switch bays, meaning, only Ethernet capable devices can be installed in these two switch bays.

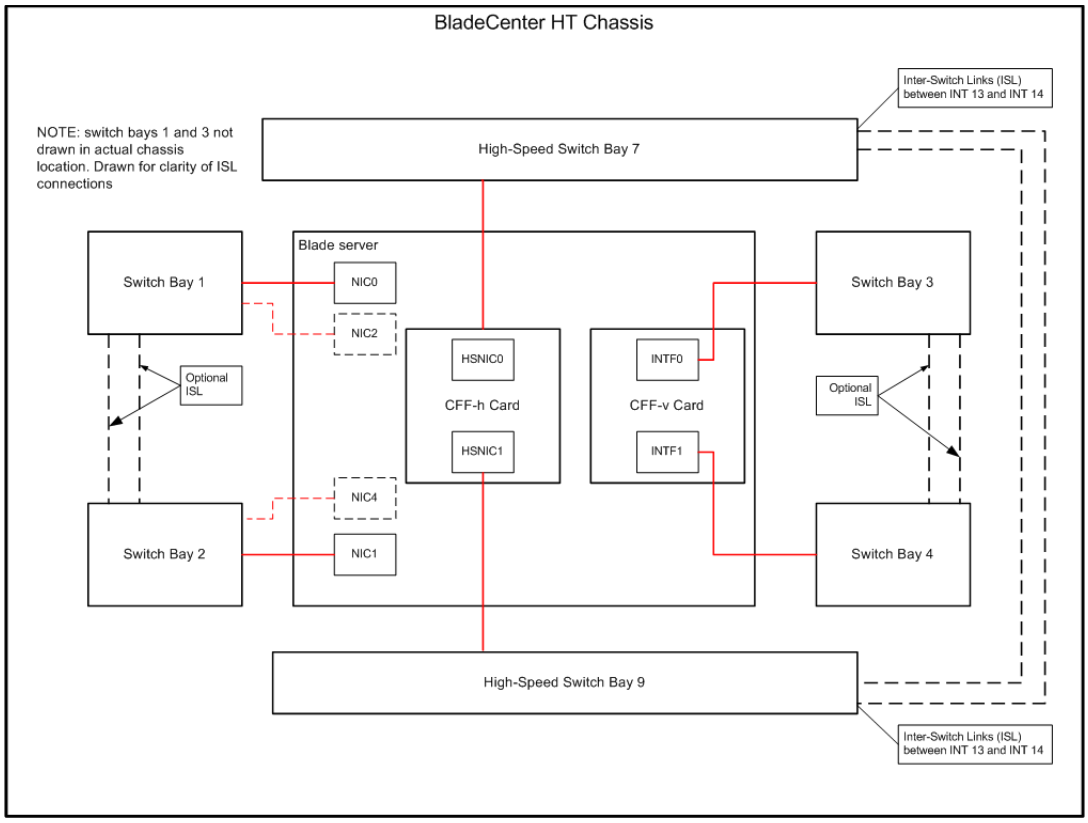

Figure 1-25 IBM BladeCenter HT Ethernet Switch Module and blade server NIC connections

Some blade servers, like the 4 Socket Double-Wide HX5 (MT7873), are 60 mm wide and occupy two server slots in the chassis. The 4 Socket Double-Wide HX5 (and other double-wide blades) have an additional pair of Ethernet interfaces on the second module of the blade. These additional NICs (depicted as NIC2 and NIC4) are also hard wired to switch bays one and two.

Blade server additional ports: 1 Gb Ethernet / Fibre Channel

Switch bays three and four are optional switch bays. These switch bays can be used for Ethernet connectivity by installing an optional CIO-v Ethernet daughter card on the blade server. Interface 0 of the daughter card (INTF0) is connected to switch bay three and interface 1 (INTF1) is connected to switch bay four. Again, these connections cannot be changed. Appropriate Ethernet switch modules must also be installed in switch bays three and four to support Ethernet connectivity to the CIO-v Ethernet card.

Blade server high speed ports: 10 Gb Ethernet / 1 Gb Ethernet / Fibre Channel

Switch bays seven to ten are optional high-speed switch bays. These switch bays are used for additional Ethernet connectivity by installing the IBM Virtual Fabric 10Gb Switch Module into the appropriate high-speed switch bays and an optional CFF-h high speed Ethernet daughter card on the blade server. If a two (or four) port daughter card is installed, then Interface 0 of the daughter card (HSNIC0) is connected to switch bay seven and interface 1 (HSNIC1) is connected to switch bay nine. If a four port card is installed in the blade server, then additional switches are required to use those additional ports. Again, these connections are hardwired from the server and cannot be changed.

Switch considerations

The IBM BladeCenter HT chassis has additional ISL links between the switch bays, as shown in Figure 1-25 on page 46. These interconnects do not change how the server's NICs interface to the switches, but allow for internal switch interconnection. The ISL interposers are optional for switch bays one and two, and three and four, but are not optional for switch bays seven to ten. Care must be taken when initially configuring switch modules where ISL interposers are involved. Network loops can be created inside the chassis if the sPanning Tree Protocol is not configured on ports 13 and 14 of the switch modules.

Ensure that you check the IBM ServerProven site for compatibility between blade servers, I/O adapters, and switches before implementation:

http://www-03.ibm.com/systems/info/x86servers/serverproven/compat/us/

More information

For more information about the IBM BladeCenter HT, see the following resources:

- IBM BladeCenter HT Announcement Letter:

  http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?infotype=AN&subtype=CA&htmlfid=897/ENUS110-209&appname=USN

- *IBM BladeCenter Products and Technology*, SG24-7523
- *IBM BladeCenter HT Installation and Users Guide*:

  http://publib.boulder.ibm.com/infocenter/bladectr/documentation/topic/com.ibm.bladecenter.8750.doc/bc_8750_iug.html

- *IBM BladeCenter HT Trouble Shooting*:

  http://publib.boulder.ibm.com/infocenter/bladectr/documentation/topic/com.ibm.bladecenter.8750.doc/bc_8750_pdsg.html

1.7 Connectors, cables, and options

In this section, we describe the most common cables and connectors that you might need to use in your implementation of the IBM 10 Gb Network infrastructure. Figure 1-26 shows the different types of connectors available for use in the IBM System Networking RackSwitches and IBM Virtual Fabric 10Gb Switch Module for IBM BladeCenter.

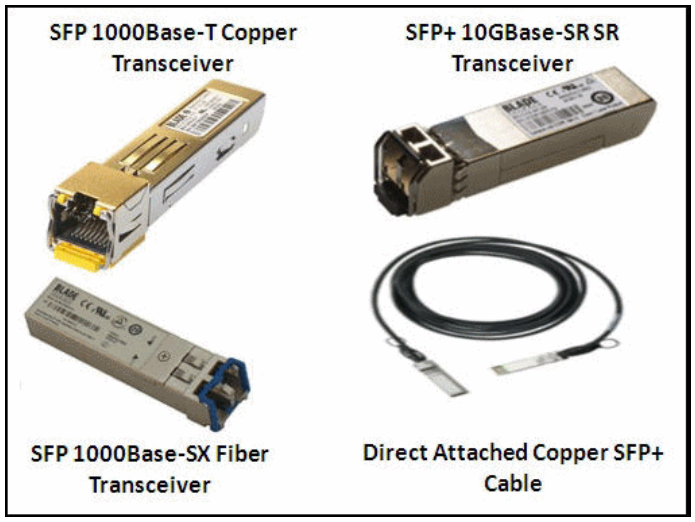
SFP 1000Base-T Copper Transceiver
SFP+ 10GBase-SR SR Transceiver
SFP 1000Base-SX Fiber Transceiver
Direct Attached Copper SFP+ Cable

Figure 1-26 Different types of SFP, SFP+, and DAC connectors for 10 Gb switching

Table 1-11 lists the part numbers and feature codes for ordering the SFP+ transceivers, FC cables, and DAC cables.

Table 1-11 IBM part numbers for ordering SFP+ transceivers, FC cables, and DAC cables

Description	Part number	Feature code
10 Gb SFP+		
IBM 10GBase-SR 10GbE 850 nm Fiber SFP+ Transceiver	44W4408	4942
IBM SFP+ Transceiver	46C3447	5053
1 Gb SFP+		
IBM 1000BASE-T (RJ45) SFP Transceiver	81Y1618	3268
IBM 1000BASE-SX SFP Transceiver	81Y1622	3269
DAC cables		
0.5 m Molex Direct Attach Copper SFP+ Cable	59Y1932	3735
1 m Molex Direct Attach Copper SFP+ Cable	59Y1936	3736
3 m Molex Direct Attach Copper SFP+ Cable	59Y1940	3737
7 m Molex Direct Attach Copper SFP+ Cable	59Y1944	3738
FC cables		
3 m Intel Connects Optical Cable	46D0153	3852

Description	Part number	Feature code
10 m Intel Connects Optical Cable	46D0156	3853
30 m Intel Connects Optical Cable	46D0159	3854

Note: The IBM Virtual Fabric 10Gb Switch Module and IBM RackSwitches can use DAC cables that are *MSA compliant*.

To assist you in selecting the different types of cables and connectors, Figure 1-27 shows the distances that each cable and connector combination can reach. It is important to understand the limitations imposed by the cabling infrastructure when planning the layout of your 10 Gb network.

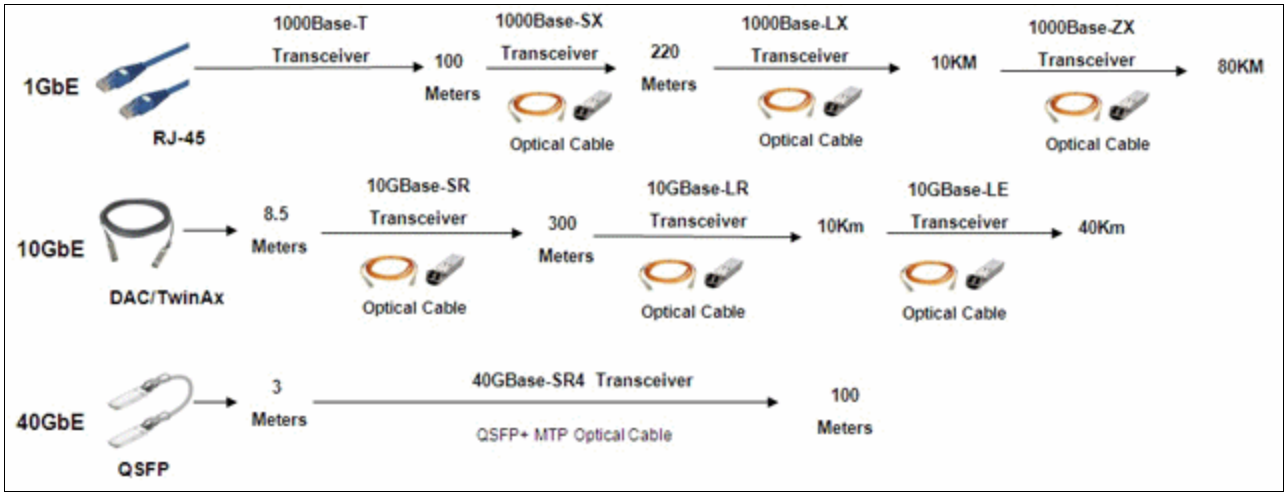

Figure 1-27 Distance guide for cables and connectors

The type of optical cable used in your implementation is important and an understanding of the limitations helps your planning. For a summary of optical cable limitations, go to the Storage Networking Industry Association (SNIA) website at:

http://www.snia.org

1.7.1 More information

For more information about cables and options, go to the IBM System Networking Options website at:

http://www-03.ibm.com/systems/networking/options/

1.8 Product interoperability

IBM and BLADE Network Technologies used The Tolly Group to certify many of its products. The Tolly Group is a leading third-party provider of validation services for vendors of IT products components and services.

BLADE Network Technologies (before being acquired by IBM) commissioned Tolly to evaluate the functionality of its RackSwitch G8000, RackSwitch G8124, and IBM Virtual Fabric 10G Switch Module for the IBM BladeCenter against a Cisco Nexus 5010 switch. Functionality tests focused on auto-negotiation, 10 GbE LAN PHY support, IEEE 802.1p/Q, Jumbo Frame support, Link Aggregation Control Protocol (LACP)/EtherChannel support, Rapid Spanning Tree Protocol (RSTP/PVRST+), Multiple Spanning Tree Protocol (MSTP) support and Fibre Channel over Ethernet (FCoE)/ Data Center Bridging (DCB) feature support. Finally, the IBM switches were tested for interoperability with the Nexus 5010 using a mixture of these features. The report can be found at:

- http://www.tolly.com/DocDetail.aspx?DocNumber=210140
- http://www.tolly.com/ts/2010/Blade/Tolly210140BLADESwitchesInteroperabilityWithCiscoNexus5010.pdf

The document is titled *BLADE Network Technologies RackSwitch G8000, RackSwitch G8124 and Virtual Fabric 10G Switch module, Functionality Certification, and Cooperative Interoperability Evaluation with Cisco Nexus 5010*.

Even though updates to the switch operating system are ongoing, this level of interoperability is maintained.

2

IBM System Networking Switch 10Gb Ethernet switch features

In Chapter 1, “Introduction to IBM System Networking 10Gb Ethernet products” on page 1, we provided an overview of the various features that are available on the different IBM System Networking 10Gb Ethernet switches. In this chapter, we describe those features in detail.

When planning a network, you must decide how the network needs to function. Your decisions should be based on the needs of your organization and typically change over time. A thought out network design leads to easy transitions as the needs of your organization changes.

This chapter provides an explanation of the various features that are available on the IBM System Networking 10Gb Ethernet switches.

© Copyright IBM Corp. 2012. All rights reserved.

2.1 Virtual Local Area Networks

This section describes network design and topology considerations for using Virtual Local Area Networks (VLANs). VLANs commonly are used to split up groups of network users into manageable broadcast domains, to create logical segmentation of workgroups, and to enforce security policies among logical segments.

2.1.1 VLANs overview

Setting up VLANs is a way to segment networks, which increases network flexibility without changing the physical network topology. With network segmentation, each switch port connects to a segment that is a single broadcast domain. When a switch port is configured to be a member of a VLAN, it is added to a group of ports (workgroup) that belong to one broadcast domain.

Ports are grouped into broadcast domains by assigning them to the same VLAN. Frames received in one VLAN can be forwarded only within that VLAN, and multicast, broadcast, and unknown unicast frames are flooded only to ports in the same VLAN.

IBM System Networking switches support jumbo frames with a Maximum Transmission Unit (MTU) of 9,216 bytes. Within each frame, 18 bytes are reserved for the Ethernet header and CRC trailer. The remaining space in the frame (up to 9,198 bytes) comprises the packet, which includes the payload of up to 9,000 bytes and any additional impact, such as 802.1q or VLAN tags. Jumbo frame support is automatic: It is enabled by default, requires no manual configuration, and cannot be manually disabled.

2.1.2 VLANs and Port VLAN ID numbers

IBM System Networking switches support up to 1024 VLANs per switch. Even though the maximum number of VLANs supported at any time is 1024, each can be identified by any number 1 - 4095.

VLAN 1 is the default VLAN for the data ports. VLAN 4095 is used by the management network, which includes the management port.

PVID numbers

Each port in the switch has a configurable default VLAN number, known as its PVID. By default, the PVID for all non-management ports is set to 1, which correlates to the default VLAN ID. The PVID for each port can be configured to any VLAN number 1 - 4094.

Each port on the switch can belong to one or more VLANs, and each VLAN can have any number of switch ports in its membership. Any port that belongs to multiple VLANs, however, must have VLAN tagging enabled.

VLAN tagging

IBM Networking OS supports 802.1Q VLAN tagging, providing standards-based VLAN support for Ethernet systems.

Tagging places the VLAN identifier in the frame header of a packet, allowing each port to belong to multiple VLANs. When you add a port to multiple VLANs, you also must enable tagging on that port.

Because tagging changes the format of frames transmitted on a tagged port, you must plan network designs to prevent tagged frames from being transmitted to devices that do not support 802.1Q VLAN tags, or devices where tagging is not enabled.

Important terms used with the 802.1Q tagging feature are:

- VLAN identifier (VID): The 12-bit portion of the VLAN tag in the frame header that identifies an explicit VLAN.
- Port VLAN identifier (PVID): A classification mechanism that associates a port with a specific VLAN. For example, a port with a PVID of 3 (PVID =3) assigns all untagged frames received on this port to VLAN 3. Any untagged frames received by the switch are classified with the PVID of the receiving port.
- Tagged frame: A frame that carries VLAN tagging information in the header. This VLAN tagging information is a 32-bit field (VLAN tag) in the frame header that identifies the frame as belonging to a specific VLAN. Untagged frames are marked (tagged) with this classification as they leave the switch through a port that is configured as a tagged port.
- Untagged frame: A frame that does not carry any VLAN tagging information in the frame header.
- Untagged member: A port that is configured as an untagged member of a specific VLAN. When an untagged frame exits the switch through an untagged member port, the frame header remains unchanged. When a tagged frame exits the switch through an untagged member port, the tag is stripped and the tagged frame is changed to an untagged frame.
- Tagged member: A port that is configured as a tagged member of a specific VLAN. When an untagged frame exits the switch through a tagged member port, the frame header is modified to include the 32-bit tag associated with the PVID. When a tagged frame exits the switch through a tagged member port, the frame header remains unchanged (the original VID remains).

Tagged frame: If a 802.1Q tagged frame is received by a port that has VLAN-tagging disabled and the port VLAN ID (PVID) is different from the VLAN ID of the packet, then the frame is dropped at the ingress port.

The default configuration settings for IBM System Networking switches have all ports set as untagged members of VLAN 1 with all ports configured as PVID = 1. In the default configuration example shown in Figure 2-1, all incoming packets are assigned to VLAN 1 by the default port VLAN identifier (PVID =1).

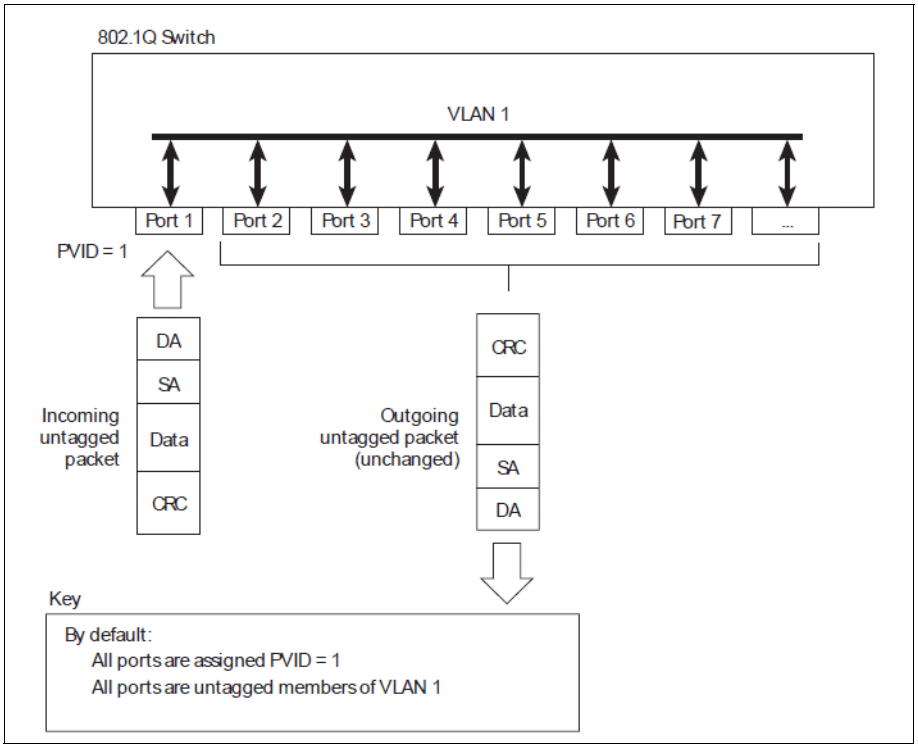

Figure 2-1 Default VLAN settings

When a VLAN is configured, ports are added as members of the VLAN, and the ports are defined as either tagged or untagged (see Figure 2-2 through Figure 2-5 on page 56).

Figure 2-2 through Figure 2-5 on page 56 show generic examples of VLAN tagging. In Figure 2-2, untagged incoming packets are assigned directly to VLAN 2 (PVID = 2). Port 5 is configured as a tagged member of VLAN 2, and port 7 is configured as an untagged member of VLAN 2.

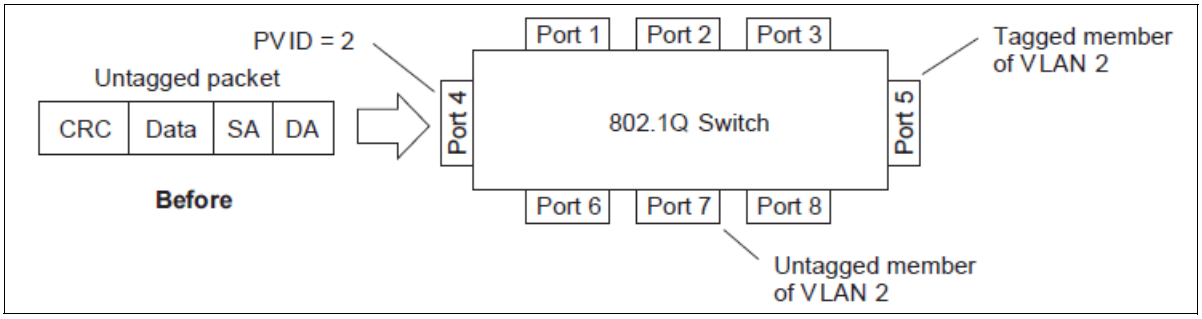

Figure 2-2 Port-based VLAN assignment

As shown in Figure 2-3, the untagged packet is marked (tagged) as it leaves the switch through port 5, which is configured as a tagged member of VLAN 2. The untagged packet remains unchanged as it leaves the switch through port 7, which is configured as an untagged member of VLAN 2.

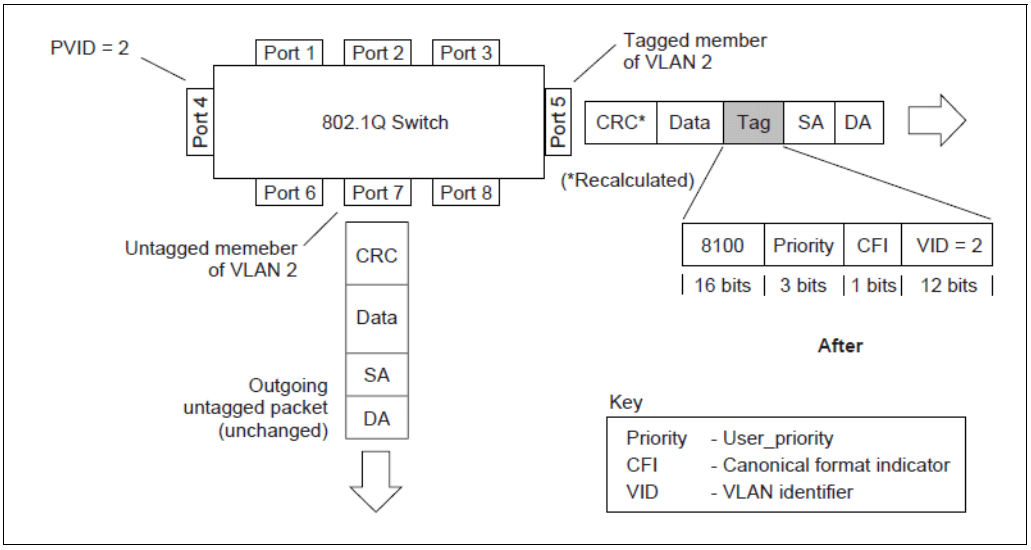

Figure 2-3 802.1Q tagging (after port-based VLAN assignment)

In Figure 2-4, tagged incoming packets are assigned directly to VLAN 2 because of the tag assignment in the packet. Port 5 is configured as a tagged member of VLAN 2, and port 7 is configured as an untagged member of VLAN 2.

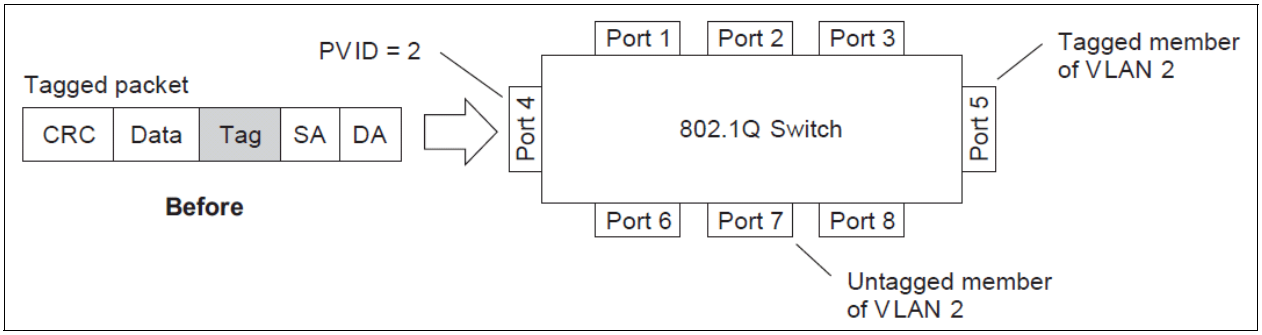

Figure 2-4 802.1q tag assignment

As shown in Figure 2-5, the tagged packet remains unchanged as it leaves the switch through port 5, which is configured as a tagged member of VLAN 2. However, the tagged packet is stripped (untagged) as it leaves the switch through port 7, which is configured as an untagged member of VLAN 2.

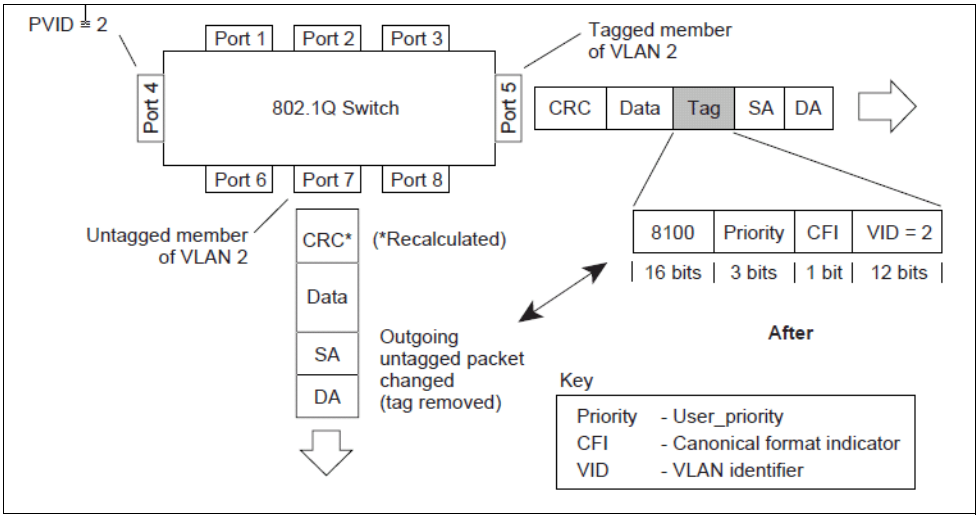

Figure 2-5 802.1Q tagging (after 802.1Q tag assignment)

2.1.3 Protocol-based VLANs

You can use protocol-based VLANs (PVLANs) to segment network traffic according to the network protocols in use. Traffic for supported network protocols can be confined to a particular port-based VLAN. You can give different priority levels to traffic generated by different network protocols.

With PVLAN, the switch classifies incoming packets by the Ethernet protocol of the packets, not by the configuration of the ingress port. When an untagged or priority-tagged frame arrives at an ingress port, the protocol information carried in the frame is used to determine the VLAN to which the frame belongs. If a frame's protocol is not recognized as a predefined PVLAN type, the ingress port's PVID is assigned to the frame. When a tagged frame arrives, the VLAN ID in the frame's tag is used.

Each VLAN can contain up to eight different PVLANs. You can configure separate PVLANs on different VLANs, with each PVLAN segmenting traffic for the same protocol type. For example, you can configure PVLAN 1 on VLAN 2 to segment IPv4 traffic, and PVLAN 8 on VLAN 100 to segment IPv4 traffic.

To define a PVLAN on a VLAN, configure a PVLAN number (1 - 8) and specify the frame type and the Ethernet type of the PVLAN protocol. You must assign at least one port to the PVLAN before it can function. Define the PVLAN frame type and Ethernet type as follows:

- Frame type: Consists of one of the following values:
  - Ether2 (Ethernet II)
  - SNAP (Subnetwork Access Protocol)
  - LLC (Logical Link Control)

- Ethernet type: Consists of a 4-digit (16 bit) hex value that defines the Ethernet type. You can use common Ethernet protocol values, or define your own values. Here are examples of common Ethernet protocol values:
  - IPv4 = 0800
  - IPv6 = 86dd
  - ARP = 0806

Port-based versus protocol-based VLANs

Each VLAN supports both port-based and protocol-based association, as follows:

- The default VLAN configuration is port-based. All data ports are members of VLAN 1, with no PVLAN association.
- When you add ports to a PVLAN, the ports become members of both the port-based VLAN and the PVLAN. For example, if you add port 1 to PVLAN 1 on VLAN 2, the port also becomes a member of VLAN 2.
- When you delete a PVLAN, its member ports remain members of the port-based VLAN. For example, if you delete PVLAN 1 from VLAN 2, port 1 remains a member of VLAN 2.
- When you delete a port from a VLAN, the port is deleted from all corresponding PVLANs.

PVLAN priority levels

You can assign each PVLAN a priority value of 0 - 7, used for Quality of Service (QoS). PVLAN priority takes precedence over a port's configured priority level. If no priority level is configured for the PVLAN (priority = 0), each port's priority is used (if configured).

All member ports of a PVLAN have the same PVLAN priority level.

PVLAN tagging

When PVLAN tagging is enabled, the switch tags frames that match the PVLAN protocol. For more information about tagging, see "VLAN tagging" on page 52.

Untagged ports must have PVLAN tagging disabled. Tagged ports can have PVLAN tagging either enabled or disabled.

PVLAN tagging has higher precedence than port-based tagging. If a port is tag enabled, and the port is a member of a PVLAN, the PVLAN tags egress frames that match the PVLAN protocol.

2.2 Spanning Tree Protocol

In high-availability environments, a redundant design is often introduced to minimize any network downtime. The redundancy is implemented on many layers, from physical cabling to redundant switches, to ensure continuous operations.

For more information about network availability protocols and technologies, see 2.7, “High availability” on page 77.

In a redundant multi-path network, Ethernet broadcast and unknown unicast flooding mechanisms can lead to forwarding loops.

In Figure 2-6, imagine a situation where Server A wants to communicate with Server B, and the MAC address of Server B is unknown.

The frame with the source MAC address of Server A’s NIC arrives at SW-1, and because the destination MAC address (Server B’s NIC MAC address) is unknown, the frame is flooded out on all links to SW-3 and SW-4. SW-3 and SW-4 and all of these switches follow the behavior of forwarding the frame out of all their interfaces and the frame finally arrives at SW-1 and is flooded again.

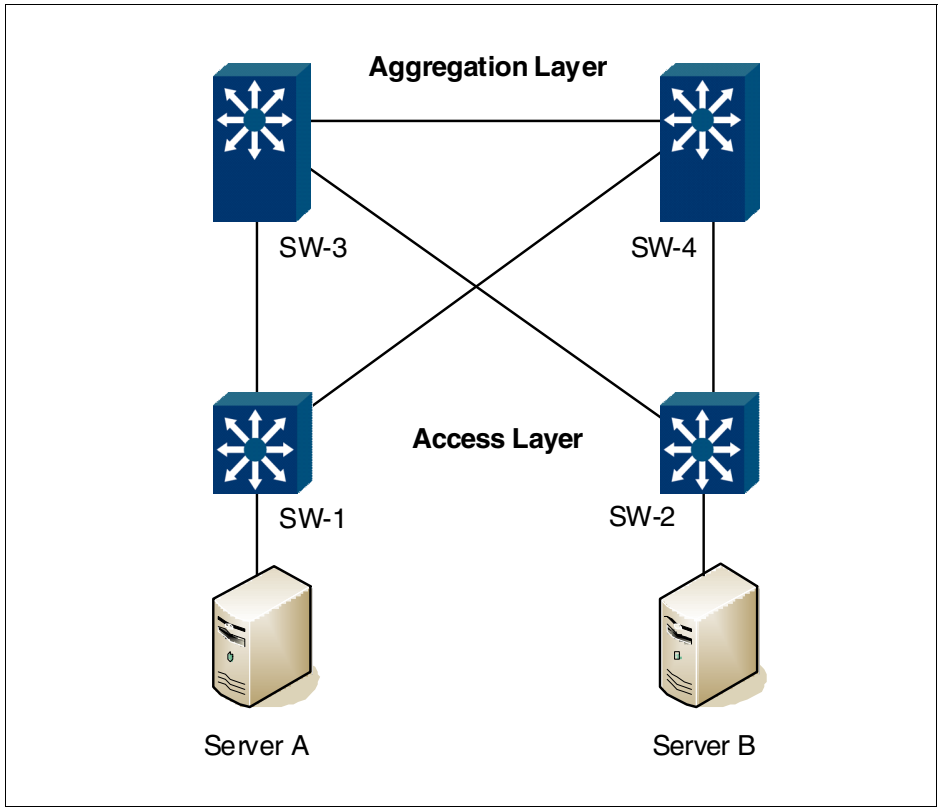

Figure 2-6 Redundant Ethernet environment

This unwanted behavior can lead to the depletion of switch resources and the network becoming non-operational. A mechanism is needed to prevent forwarding loops in a switched Ethernet environment. The mechanism is part of the IEEE 802.1d standard and is known as the Spanning Tree Protocol (STP).

The details of the operations of STP are not covered in this book. What is important to remember is that STP works in Layer 2 by detecting forwarding loops and logically disabling the link that is part of the loop. STP operates by transmitting and receiving Bridge Protocol Data Units (BPDUs).

For more information about STP, see the originalIEEE 802.1d specification at the following website:

http://standards.ieee.org/getieee802/download/802.1D-2004.pdf

There are multiple flavors of STP, as described in the following sections.

2.2.1 Rapid Spanning Tree Protocol

Rapid Spanning Tree Protocol (RSTP) (802.1w) provides rapid convergence of the Spanning Tree and provides the fast reconfiguration critical for networks that carry delay-sensitive traffic, such as voice and video. RSTP significantly reduces the time to reconfigure the active topology of the network when changes occur to the physical topology or its configuration parameters.

RSTP is compatible with devices that run IEEE 802.1d STP. If the switch detects IEEE 802.1d BPDUs, it responds with IEEE 802.1d-compatible data units.

2.2.2 Per-VLAN Rapid Spanning Tree Protocol (PVRST)

Per-VLAN Rapid Spanning Tree Protocol (PVRST) is based on IEEE 802.1w RSTP. Like RSTP, PVRST mode provides rapid Spanning Tree convergence. However, similar to the way standard STP allows per-VLAN spanning-tree instance on a switch. Each VLAN has its own Spanning Tree instance and tree, so that the VLANs can use different paths. The drawback is that the number of Spanning Tree instances can grow significantly and affect processing time on the switches in networks with many VLANs.

2.2.3 Multiple Spanning Tree Protocol

Multiple Spanning Tree Protocol (MSTP) extends Rapid Spanning Tree Protocol (RSTP), allowing multiple Spanning Tree instances, which may each include multiple VLANs.

In MSTP mode, IBM System Networking switches support up to 32 instances of Spanning Tree. MSTP allows frames assigned to different VLANs to follow separate paths, with each path based on an independent Spanning Tree instance. This approach provides multiple forwarding paths for data traffic, enabling load-balancing, and reducing the number of Spanning Tree instances required to support many VLANs.

2.3 IP routing

IBM System Networking switches use a combination of configurable IP switch interfaces and IP routing options. The switch IP routing capabilities provide the following benefits:

- Connects the server IP subnets to the rest of the backbone network.
- Provides routing of IP traffic between multiple VLANs configured on the switch.

The physical layout of most corporate networks has evolved over time. Classic hub/router topologies have given way to faster switched topologies, particularly now that switches are increasingly intelligent. IBM System Networking switches are intelligent and fast enough to perform routing functions on a par with wirespeed Layer 2 switching. The combination of faster routing and switching in a single device provides another service: You can build versatile topologies that account for earlier configurations.

Figure 2-7 shows a corporate campus migrated from a router-centric topology to a faster, more powerful, switch-based topology. As often the case is, network growth and redesign has left the system with a mix of illogically distributed subnets.

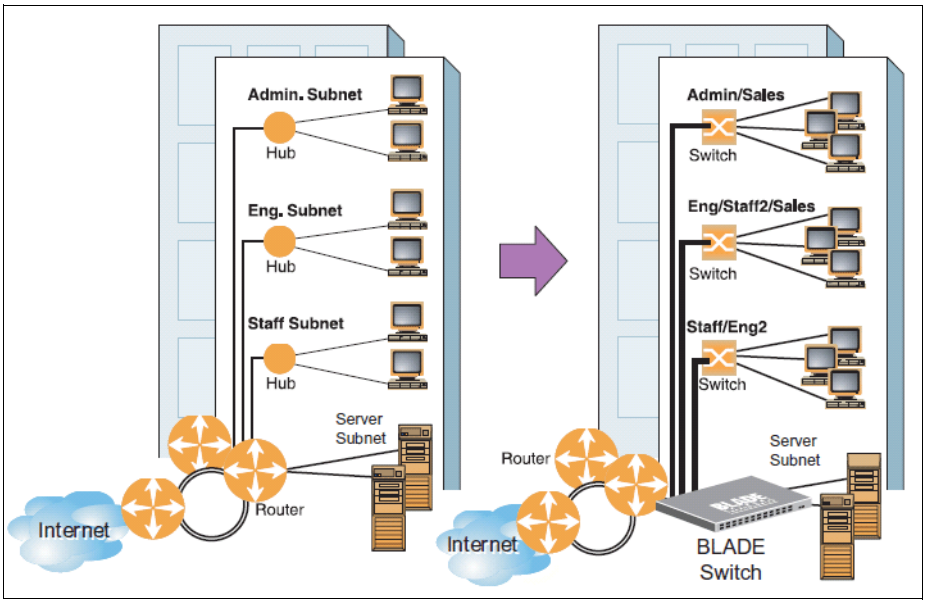

Figure 2-7 Router-centric versus switch-based network topologies

This situation in one that switching alone cannot cure. Instead, the router is flooded with cross-subnet communication. This compromises efficiency in two ways:

- Routers can be slower than switches. The cross-subnet side trip from the switch to the router and back again adds two hops for the data, slowing throughput considerably.
- Traffic to the router increases, increasing congestion.

Even if every endpoint could be moved to better logical subnets (a daunting task), competition for access to common server pools on different subnets still burdens the routers.

This problem is solved by using switches with built-in IP routing capabilities. Because IBM System Networking switches use ASIC for forwarding Layer 3 packets, cross-subnet LAN traffic can now be routed within the switches with wirespeed Layer 2 switching performance. This configuration eases not only the load on the router, but saves the network administrators from reconfiguring each endpoint with new IP addresses.

2.3.1 Static routes

You can use static routes to forward an IP packet based on a manually configured entry. The entry specifies a network and an IP address of a gateway, or next-hop, to that network.

2.3.2 Equal-Cost Multi-Path static routes

Equal-Cost Multi-Path (ECMP) is a forwarding mechanism that routes packets along multiple paths of equal cost. ECMP provides equally distributed link load sharing across the paths. The hashing algorithm used is based on the source IP address. ECMP routes allow the switch to choose between several next hops toward a destination. The switch performs periodic health checks (ping) on each ECMP gateway. If a gateway fails, it is removed from the routing table, and an SNMP trap is sent.

2.3.3 Routing Information Protocol

In a routed environment, routers communicate with one another to track available routes. Routers can learn about available routes dynamically by using the Routing Information Protocol (RIP). IBM Networking OS supports RIP version 1 (RIPv1) and RIP version 2 (RIPv2) for exchanging TCP/IPv4 route information with other routers.

Distance vector protocol

RIP is known as a distance vector protocol. The vector is the network number and next hop, and the distance is the metric associated with the network number. RIP identifies network reachability based on a metric, and the metric is defined as a hop count. One hop is considered to be the distance from one router to the next, which typically is 1. When a router receives a routing update that contains a new or changed destination network entry, the router adds 1 to the metric value indicated in the update and enters the network in the routing table. The IPv4 address of the sender is used as the next hop.

Stability

RIP includes a number of other stability features that are common to many routing protocols. For example, RIP implements the split horizon and hold-down mechanisms to prevent incorrect routing information from being propagated.

RIP prevents routing loops from continuing indefinitely by implementing a limit on the number of hops allowed in a path from the source to a destination. The maximum number of hops in a path is 15. The network destination network is considered unreachable if increasing the metric value by 1 causes the metric to be 16 (that is, infinity). This setting limits the maximum diameter of a RIP network to less than 16 hops.

RIP is often used in stub networks and in small autonomous systems that do not have many redundant paths.

Routing updates

RIP sends routing-update messages at regular intervals and when the network topology changes. Each router “advertises” routing information by sending a routing information update every 30 seconds. If a router does not receive an update from another router for 180 seconds, those routes provided by that router are declared invalid. The routes are removed from the routing table, but they remain in the RIP routes table. After another 120 seconds without receiving an update for those routes, the routes are removed from respective regular updates.

When a router receives a routing update that includes changes to an entry, it updates its routing table to reflect the new route. The metric value for the path is increased by 1, and the sender is indicated as the next hop. RIP routers maintain only the best route (the route with the lowest metric value) to a destination.

RIPv1

RIP version 1 use broadcast User Datagram Protocol (UDP) data packets for the regular routing updates. The main disadvantage is that the routing updates do not carry subnet mask information. Hence, the router cannot determine whether the route is a subnet route or a host route. RIPv1 is of limited usage after the introduction of RIPv2. For more information about RIPv1 and RIPv2, see RFC 1058, found at:

http://www.ietf.org/rfc/rfc1058.txt

RIPv2

RIPv2 is the most popular and preferred configuration for most networks. RIPv2 expands the amount of useful information carried in RIP messages and provides a measure of security. For a detailed explanation of RIPv2, see the following RFCs:

- RFC 1723, found at:

  http://www.ietf.org/rfc/rfc1723.txt

- RFC 2453, found at:

  http://www.ietf.org/rfc/rfc2453.txt

RIPv2 improves efficiency by using multicast UDP (address 224.0.0.9) data packets for regular routing updates. Subnet mask information is provided in the routing updates. A security option is added for authenticating routing updates, by using a shared password.

2.3.4 Open Shortest Path First

Open Shortest Path First (OSPF) is designed for routing traffic within a single IP domain called an Autonomous System (AS). The AS can be divided into smaller logical units known as *areas*.

All routing devices maintain link information in their own Link State Database (LSDB). The LSDB for all routing devices within an area is identical, but is not exchanged between different areas. Only routing updates are exchanged between areas, reducing the impact for maintaining routing information on a large, dynamic network.

OSPF area types

An AS can be broken into logical units known as *areas*. In any AS with multiple areas, one area must be designated as area 0, known as the *backbone*. The backbone acts as the central OSPF area.

All other areas in the AS must be connected to the backbone. Areas inject summary routing information into the backbone, which then distributes it to other areas as needed.

OSPF defines the following types of areas (shown in Figure 2-8):

- Stub area: An area that is connected to only one other area. External route information is not distributed into stub areas.
- Not-So-Stubby-Area (NSSA): Similar to a stub area with additional capabilities. Routes originating from within the NSSA can be propagated to adjacent transit and backbone areas. External routes from outside the AS can be advertised within the NSSA but are not distributed into other areas.
- Transit Area: An area that allows area summary information to be exchanged between routing devices. The backbone (area 0), any area that contains a virtual link to connect two areas, and any area that is not a stub area or an NSSA are considered transit areas.

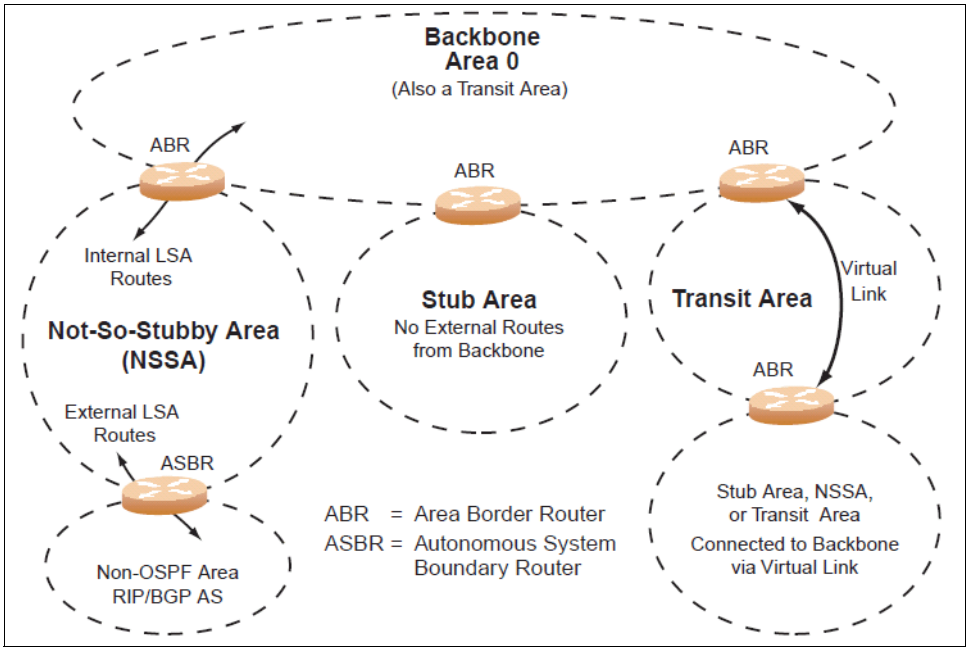

Figure 2-8 OSPF area types

OSPF router types

As shown in Figure 2-9 OSPF uses the following types of routing devices:

- Internal Router (IR): A router that has all of its interfaces within the same area. IRs maintain LSDBs identical to the LSDBs of other routing devices within the local area.
- Area Border Router (ABR): A router that has interfaces in multiple areas. ABRs maintain one LSDB for each connected area and disseminate routing information between areas.
- Autonomous System Boundary Router (ASBR): A router that acts as a gateway between the OSPF domain and non-OSPF domains, such as RIP, BGP, and static routes.

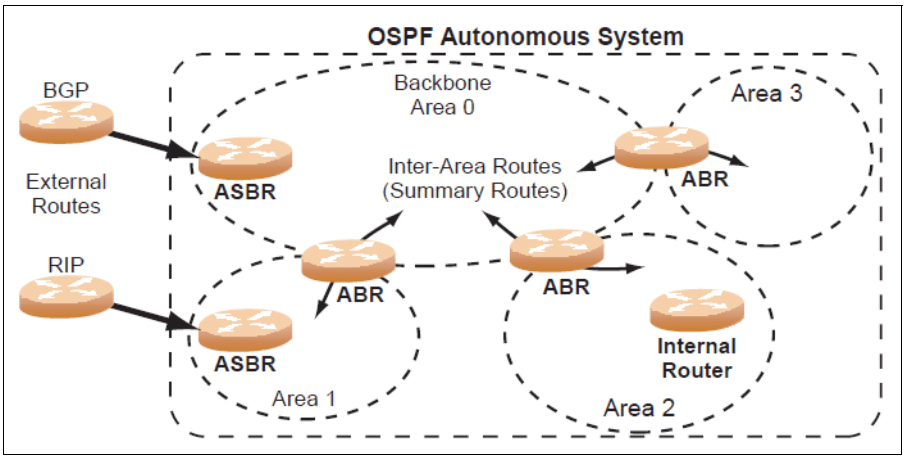

Figure 2-9 OSPF router types

Neighbors and adjacencies

In areas with two or more routing devices, *neighbors* and *adjacencies* are formed. Neighbors are routing devices that maintain information about each others' health. To establish neighbor relationships, routing devices periodically send hello packets on each of their interfaces. All routing devices that share a common network segment, appear in the same area, and have the same health parameters (*hello* and *dead* intervals), authentication parameters, area number, and area stub-flag respond to each other's hello packets and become neighbors.

Neighbors continue to send periodic hello packets to advertise their health to neighbors. In turn, they listen to hello packets to determine the health of their neighbors and to establish contact with new neighbors. On broadcast networks (like Ethernet), the hello process is used for electing one of the neighbors as the area's Designated Router (DR) and one as the area's Backup Designated Router (BDR). The DR is next to all other neighbors and acts as the central contact for database exchanges. Each neighbor sends its database information to the DR, which relays the information to the other neighbors.

The BDR is next to all other neighbors (including the DR). Each neighbor sends its database information to the BDR as with the DR, but the BDR merely stores this data and does not distribute it. If the DR fails, the BDR takes over the task of distributing database information to the other neighbors.

Link State Database

OSPF is a link-state routing protocol. A *link* represents an interface (or routable path) from the routing device. By establishing an adjacency with the DR, each routing device in an OSPF area maintains an identical Link-State Database (LSDB) describing the network topology for its area.

Each routing device transmits a Link-State Advertisement (LSA) on each of its *active* interfaces. LSAs are entered into the LSDB of each routing device. OSPF uses flooding to distribute LSAs between routing devices. Interfaces may also be *passive*. Passive interfaces send LSAs to active interfaces, but do not receive LSAs, hello packets, or any other OSPF protocol information from active interfaces. Passive interfaces behave as stub networks, allowing OSPF routing devices to be aware of devices that otherwise participate in OSPF (either because they do not support it, or because the administrator chooses to restrict OSPF traffic exchange or transit).

When LSAs result in changes to the routing device's LSDB, the routing device forwards the changes to the adjacent neighbors (the DR and BDR) for distribution to the other neighbors.

OSPF routing updates occur only when changes occur, instead of periodically. For each new route, if an adjacent neighbor is interested in that route, an update message that contains the new route is sent to the neighbor. For each route removed from the route table, if the route has already been sent to an adjacent neighbor, an update message that contains the route to withdraw is sent.

Shortest Path First

The routing devices use a link-state algorithm (Dijkstra's algorithm) to calculate the shortest path to all known destinations, based on the cumulative cost required to reach the destination.

The cost of an individual interface in OSPF is an indication of the processing required to send packets across it. The cost is inversely proportional to the bandwidth of the interface. A lower cost indicates a higher bandwidth.

Internal versus external routing

To ensure effective processing of network traffic, every routing device on your network needs to know how to send a packet (directly or indirectly) to any other location/destination in your network. This action is referred to as *internal routing* and can be done with static routes or using active internal routing protocols, such as OSPF or RIP.

It is also useful to tell routers outside your network (upstream providers or peers) about the routes you have access to in your network. Sharing of routing information between autonomous systems is known as *external routing*.

Typically, an AS has one or more border routers (peer routers that exchange routes with other OSPF networks), and an internal routing system that enables every router in that AS to reach every other router and destination within that AS.

When a routing device advertises routes to boundary routers on other autonomous systems, it is committing to carry data to the IP space represented in the route that is advertised. For example, if the routing device advertises `192.204.4.0/24`, it is declaring that if another router sends data destined for any address in the `192.204.4.0/24` range, it carries that data to its destination.

2.3.5 Border Gateway Protocol

Border Gateway Protocol (BGP) is an Internet protocol that enables routers on an IPv4 network to share and advertise routing information with each other about the segments of the IPv4 address space they can access within their network and with routers on external networks. You can use BGP to decide what is the “best” route for a packet to take from your network to a destination on another network rather than setting a default route from your border routers to your upstream providers. BGP is defined in RFC 1771, found at the following website:

`http://www.ietf.org/rfc/rfc1771.txt`

External networks (outside your own) that are under the same administrative control are referred to as *autonomous systems* (AS). Sharing of routing information between autonomous systems is known as external routing.

External BGP (eBGP) is used to exchange routes between different autonomous systems, and internal BGP (iBGP) is used to exchange routes within the same autonomous system. An iBGP is a type of internal routing protocol you can use to do active routing inside your network. It also carries AS path information, which is important when you are an ISP or doing BGP transit.

The iBGP peers must maintain reciprocal sessions to every other iBGP router in the same AS (in a full-mesh manner) to propagate route information throughout the AS.

If the iBGP session shown between the two routers in AS 20 is not present (Figure 2-10), the top router does not learn the route to AS 50, and the bottom router does not learn the route to AS 11, even though the two AS 20 routers are connected through the IBM System Networking switch.

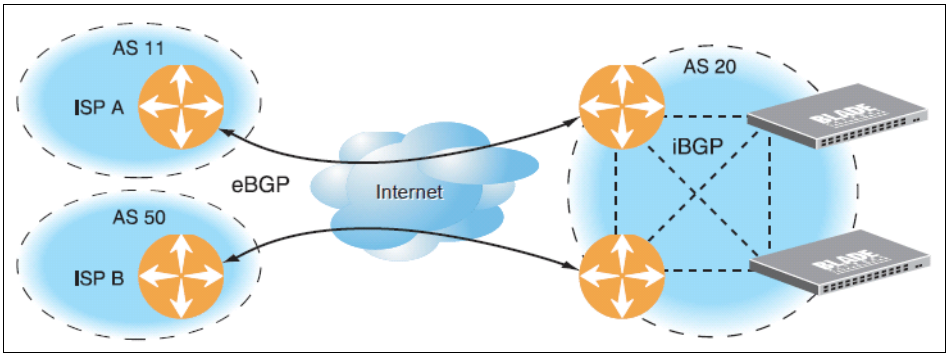

Figure 2-10 iBGP and eBGP

Typically, an AS has one or more *border routers*, which are peer routers that exchange routes with other ASs, and an internal routing scheme that enables routers in that AS to reach every other router and destination within that AS. When you advertise routes to border routers on other autonomous systems, you are committing to carry data to the IPv4 space represented in the route that is advertised. For example, if you advertise `192.204.4.0/24`, you are declaring that if another router sends you data destined for any address in `192.204.4.0/24`, you know how to carry that data to its destination.

Forming BGP peer routers

Two BGP routers become peers or neighbors after you establish a TCP connection between them. For each new route, if a peer is interested in that route (if a peer would like to receive your static routes and the new route is static), an update message is sent to the peer that contains the new route. For each route removed from the route table, if the route has already been sent to a peer, an update message that contains the route to withdraw is sent to that peer.

For each Internet host, you must be able to send a packet to that host, and that host must have a path back to you. This setup means that whoever provides Internet connectivity to that host must have a path to you. Ultimately, this means that they must “hear a route” that covers the section of the IPv4 space you are using; otherwise, you do not have connectivity to the host in question.

2.4 IP multicast

IP multicast represents a one-to-many communications scheme with one host that sends a stream and a number of hosts that receive it.

IPv4 multicast always uses UDP as a transport layer protocol with class D (224.0.0.0 - 239.255.255.255) IP addresses as the Layer 3 destination address.

There are two major protocol types used in a multicast network:

- Internet Group Management Protocol (IGMP): Used for hosts that signal a router that they are interested in receiving multicast traffic destined for the particular group (or multicast address). IGMP is a “host-to-router” multicast protocol.
- Protocol Independent Multicast (PIM): For providing loop-free multicast traffic delivery in a network. PIM is considered a “router-to-router” multicast protocol.

For more information, see *IP Multicast Protocol Configuration*, REDP-4777.

2.4.1 Internet Group Management Protocol

Internet Group Management Protocol (IGMP) is used by IPv4 Multicast routers to learn about the existence of host group members on their directly attached subnet (for more information, see RFC 2236 at http://www.ietf.org/rfc/rfc2236.txt). The IPv4 Multicast routers get this information by broadcasting IGMP membership queries and listening for IPv4 hosts reporting their host group memberships. This process is used to set up a client/server relationship between an IPv4 Multicast router that provides the data streams on behalf of the multicast sender and the clients that want to receive the data.

IBM System Networking switches can perform IGMP Snooping, and connect to static multicast routers (Mrouters). They can act asan IGMPQuerier, and participate in the IGMP Querier election process.

IBM System Networking switches support IGMP versions 1, 2, and 3.

IGMP Snooping

IGMP Snooping allows the switch to forward multicast traffic only to those ports that request it. IGMP Snooping prevents multicast traffic from being flooded to all ports. The switch learns which server hosts are interested in receiving multicast traffic, and forwards it only to ports connected to those servers.

IGMP Snooping conserves bandwidth. With IGMP Snooping, the switch learns which ports are interested in receiving multicast data, and forwards multicast data only to those ports. In this way, other ports are not burdened with unwanted multicast traffic.

The switch can sense IGMP Membership Reports from attached clients and act as a proxy to set up a dedicated path between the requesting host and a local IPv4 Multicast router. After the pathway is established, the switch blocks the IPv4 Multicast stream from flowing through any port that does not connect to a host member, thus conserving bandwidth.

The client-server path is set up as follows:

1. An IPv4 Multicast Router (Mrouter) sends Membership Queries to the switch, which forwards them to all ports in a VLAN.
2. Hosts that want to receive the multicast data stream send Membership Reports to the switch, which sends a proxy Membership Report to the Mrouter.
3. The switch sets up a path between the Mrouter and the host, and blocks all other ports from receiving the multicast.
4. Periodically, the Mrouter sends Membership Queries to ensure that the host wants to continue receiving the multicast. If a host fails to respond with a Membership Report, the Mrouter stops sending the multicast to that path.
5. The host can send a Leave Report to the switch, which sends a proxy Leave Report to the Mrouter. The multicast path is terminated immediately.

IGMP entries

An IGMP entry is allocated for each unique join request, based on the VLAN and IGMP group address. If multiple ports join the same IGMP group using the same VLAN, only a single IGMP entry is used.

FastLeave

In a normal IGMP operation, when the switch receives an IGMPv2 leave message, it sends a Group-Specific Query to determine if any devices in the same group (and on the same port) are interested in the specified multicast group traffic. The switch removes the affiliated port from that particular group, if the following conditions apply:

- If the switch does not receive an IGMP Membership Report within the *query-response-interval*.
- If no multicast routers are learned on that port.

With FastLeave enabled on the VLAN, a port can be removed immediately from the port list of the group entry when the IGMP Leave message is received, unless a multicast router was learned on the port.

IGMP v3 snooping

IGMPv3 includes new membership report messages to extend IGMP functionality. The switch provides snooping capability for all types of IGMP version 3 (IGMPv3) Membership Reports. IGMPv3 supports Source-Specific Multicast (SSM). SSM identifies session traffic by both source and group addresses.

The IGMPv3 implementation keeps records on the multicast hosts present in the network. If a host is already registered, when it receives a new report from same host, the switch makes the correct transition to new (port-host-group) registration based on the IGMPv3 RFC. The registrations of other hosts for the same group on the same port are not changed.

IGMP Querier

IGMP Querier allows the switch to perform the multicast router (Mrouter) role and provide Mrouter discovery when the network or virtual LAN (VLAN) does not have a router.

When IGMP Querier is enabled on a VLAN, the switch acts as an IGMP querier in a Layer 2 network environment. The IGMP querier periodically broadcasts IGMP Queries and listens for hosts to respond with IGMP Reports indicating their IGMP group memberships. If multiple Mrouters exist on a network, the Mrouters elect one as the querier, which performs all periodic membership queries. The election process can be based on IPv4 address or MAC address.

IGMP Relay

IBM System Networking switch can act as an IGMP Relay (or IGMP Proxy) device that relays IGMP multicast messages and traffic between a Mrouter and endpoints. IGMP Relay allows a switch to participate in network multicasts with no configuration of the various multicast routing protocols, so you can deploy it in the network with minimal effort.

To an IGMP host connected to IBM System Networking switch, IGMP Relay appears to be an IGMP multicast router (Mrouter). IGMP Relay sends Membership Queries to hosts, which respond by sending an IGMP response message. A host can also send an unsolicited Join message to the IGMP Relay.

To a multicast router, IGMP Relay appears as a host. The Mrouter sends IGMP host queries to IGMP Relay, and IGMP Relay responds by forwarding IGMP host reports and unsolicited join messages from its attached hosts.

IGMP Relay also forwards multicast traffic between the Mrouter and endpoints, similar to IGMP Snooping.

You can configure up to two Mrouters to use with IGMP Relay. One Mrouter acts as the primary Mrouter, and one is the backup Mrouter. The switch uses health checks to select the primary Mrouter.

IGMP Filtering

With IGMP Filtering, you can allow or deny a port to send and receive multicast traffic to certain multicast groups. Unauthorized users are restricted from streaming multicast traffic across the network.

If access to a multicast group is denied, IGMP Membership Reports from the port are dropped, and the port is not allowed to receive IPv4 multicast traffic from that group. If access to the multicast group is allowed, Membership Reports from the port are forwarded for normal processing.

To configure IGMP Filtering, you must globally enable IGMP filtering, define an IGMP filter, assign the filter to a port, and enable IGMP Filtering on the port. To define an IGMP filter, you must configure a range of IPv4 Multicast groups, choose whether the filter allows or denies multicast traffic for groups within the range, and enable the filter.

2.4.2 Protocol Independent Multicast

Protocol Independent Multicast (PIM) is designed for efficiently routing multicast traffic across one or more IPv4 domains. This protocol has benefits for application such as IP television, collaboration, education, and software delivery, where a single source must deliver content (a multicast) to a group of receivers that span both wide-area and inter-domain networks.

Instead of sending a separate copy of content to each receiver, a multicast derives efficiency by sending only a single copy of content toward its intended receivers. This single copy becomes duplicated only when it reaches the target domain that includes multiple receivers, or when it reaches a necessary bifurcation point leading to different receiver domains.

PIM is used by multicast source stations, client receivers, and intermediary routers and switches, to build and maintain efficient multicast routing trees. PIM is protocol independent; it collects routing information by using the existing unicast routing functions underlying the IPv4 network, but does not rely on any particular unicast protocol. For PIM to function, a Layer 3 routing protocol (such as BGP, OSPF, RIP, or static routes) must first be configured on the switch.

PIM-SM is a reverse-path routing mechanism. Client receiver stations advertise their willingness to join a multicast group. The local routing and switching devices collect multicast routing information and forward the request toward the station that provides the multicast content. When the join requests reach the sending station, the multicast data is sent toward the receivers, flowing in the opposite direction of the original join requests.

Some routing and switching devices perform special PIM-SM functions. Within each receiver segment, one router is elected as the Designated Router (DR) for handling multicasts for the segment. DRs forward information to a similar device, the Rendezvous Point (RP), which holds the root tree for the particular multicast group.

Receiver join requests and sender multicast content initially converge at the RP, which generates and distributes multicast routing data for the DRs along the delivery path. As the multicast content flows, DRs use the routing tree information obtained from the RP to optimize the paths both to and from send and receive stations, bypassing the RP for the remainder of content transactions if a more efficient route is available.

DRs continue to share routing information with the RP, modifying the multicast routing tree when new receivers join, or pruning the tree when all the receivers in any particular domain are no longer part of the multicast group.

Supported PIM modes and features

For each interface attached to a PIM network component, PIM can be configured to operate either in PIM Sparse Mode (PIM-SM) or PIM Dense Mode (PIM-DM).

- PIM-SM is used in networks where multicast senders and receivers comprise a relatively small (sparse) portion of the overall network. PIM-SM uses a more complex process than PIM-DM for collecting and optimizing multicast routes, but minimizes impact on other IP services and is more commonly used.
- PIM-DM is used where multicast devices are a relatively large (dense) portion of the network, with frequent (or constant) multicast traffic. PIM-DM requires less configuration on the switch than PIM-SM, but uses broadcasts that can consume more bandwidth in establishing and optimizing routes.

PIM Dense Mode

PIM Dense Mode, which is not commonly used today, is intended for *densly populated* networks with many receivers. The PIM Dense Mode uses *Flood-Prune* behavior, which means traffic is by default flooded to each network segment. It is up to the router on the segment to send a prune message to the source, which signals that there are no receivers interested in the multicast traffic flooded.

For more information about PIM Dense Mode, see RFC 3973, found at:

http://www.ietf.org/rfc/rfc3973.txt

PIM Sparse Mode

The behavior of PIM Sparse Mode is opposite of Dense Mode. The default behavior is to not to flood the multicast traffic unless the downstream routers signal, by sending a PIM Join message, that there are receivers on their directly connected networks interested in receiving the multicast traffic.

For more information about PIM Sparse Mode, see RFC 4601, found at:

http://www.ietf.org/rfc/rfc4601.txt

2.5 IPv6

Internet Protocol version 6 (IPv6) is a network layer protocol intended to expand the network address space. IPv6 is a robust and expandable protocol that meets the need for increased physical address space.

For more information about the IPv6, see *IPv6 Introduction and Configuration*, REDP-4776.

2.5.1 IPv6 address format

The IPv6 address is 128 bits (16 bytes) long and is represented as a sequence of eight 16-bit hex values, separated by colons.

Each IPv6 address has two parts:

- A subnet prefix that represents the network to which the interface is connected.
- A local identifier, which is either derived from the MAC address or user-configured.

The preferred hexadecimal format is as follows:

```
xxxx:xxxx:xxxx:xxxx:xxxx:xxxx:xxxx:xxxx
```

An example IPv6 address is:

```
FEDC:BA98:7654:BA98:FEDC:1234:ABCD:5412
```

Some addresses can contain long sequences of zeros. A single contiguous sequence of zeros can be compressed to :: (two colons). For example, consider the following IPv6 address:

```
FE80:0:0:0:2AA:FF:FA:4CA2
```

The address can be compressed as follows:

```
FE80::2AA:FF:FA:4CA2
```

Unlike IPv4, a subnet mask is not used for IPv6 addresses. IPv6 uses the subnet prefix as the network identifier. The prefix is the part of the address that indicates the bits that have fixed values or are the bits of the subnet prefix. An IPv6 prefix is written in address/prefix-length notation. For example, in the following address, 64 is the network prefix:

```
21DA:D300:0000:2F3C::/64
```

IPv6 addresses can be either user-configured or automatically configured. Automatically configured addresses always have a 64-bit subnet prefix and a 64-bit interface identifier. In most implementations, the interface identifier is derived from the switch's MAC address, using a method called EUI-64.

Most IBM Networking OS features permit IP addresses to be configured using either IPv4 or IPv6 address formats. Throughout this manual, IP address is used in places where either an IPv4 or IPv6 address is allowed. In places where only one type of address is allowed, the type (IPv4 or IPv6) is specified.

2.5.2 IPv6 address types

IPv6 supports three types of addresses: unicast (one-to-one), multicast (one-to-many), and anycast (one-to-nearest). Multicast addresses replace the use of broadcast addresses.

Unicast addresses

Unicast is a communication between a single host and a single receiver. Packets sent to a unicast address are delivered to the interface identified by that address. IPv6 defines the following types of unicast addresses:

- Global Unicast address: An address that can be reached and identified globally. Global Unicast addresses use the high-order bit range up to FF00, therefore all non-multicast and non-link-local addresses are considered to be global unicast. A manually configured IPv6 address must be fully specified. Auto-configured IPv6 addresses are composed of a prefix combined with the 64-bit EUI. RFC 4291 (found at http://www.ietf.org/rfc/rfc4291.txt) defines the IPv6 addressing architecture.

  The interface ID must be unique within the same subnet.

- Link-local unicast address: An address used to communicate with a neighbor on the same link. Link-local addresses use the format `FE80::EUI`

  Link-local addresses are used for addressing on a single link for purposes, such as automatic address configuration, neighbor discovery, or when no routers are present.

  Routers must not forward any packets with link-local source or destination addresses to other links.

- Unique Local IPv6 Unicast addressees are synonymous to private addresses in IPv4 and are in the `FC00::/7` range.

Multicast addresses

Multicast is communication between a single host and multiple receivers. Packets are sent to all interfaces identified by that address. An interface may belong to any number of multicast groups.

A multicast address (FF00 - FFFF) is an identifier for a group interface. The multicast address most often encountered is a solicited-node multicast address using prefix `FF02::1:FF00:0000/104` with the low-order 24 bits of the unicast or anycast address.

The following well-known multicast addresses are predefined. The group IDs defined in this section are defined for explicit scope values, as follows:

`FF00:::::::0` through `FF0F:::::::0`

Anycast

Packets sent to an anycast address or list of addresses are delivered to the nearest interface identified by that address. Anycast is a communication between a single sender and a list of addresses.

Anycast addresses are allocated from the unicast address space, using any of the defined unicast address formats. Thus, anycast addresses are syntactically indistinguishable from unicast addresses. When a unicast address is assigned to more than one interface, thus turning it into an anycast address, the nodes to which the address is assigned must be configured to know that it is an anycast address.

2.5.3 IPv6 address auto-configuration

IPv6 supports the following types of address autoconfiguration:

- Stateful address configuration

  Address configuration is based on the use of a stateful address configuration protocol, such as DHCPv6, to obtain addresses and other configuration options.

- Stateless address configuration

  Address configuration is based on the receipt of Router Advertisement messages that contain one or more Prefix Information options.

IBM System Networking switches support stateless address configuration. Stateless address configuration allows hosts on a link to configure themselves with link-local addresses and with addresses derived from prefixes advertised by local routers. Even if no router is present, hosts on the same link can configure themselves with link-local addresses and communicate without manual configuration

2.5.4 Neighbor Discovery protocol

The switch uses Neighbor Discovery protocol (ND) to gather information about other router and host nodes, including the IPv6 addresses. Host nodes use ND to configure their interfaces and perform health detection. ND allows each node to determine the link-layer addresses of neighboring nodes, and to track each neighbor's information. A neighboring node is a host or a router that is linked directly to the switch. The switch supports Neighbor Discovery, as described in RFC 4861 (http://www.ietf.org/rfc/rfc4861.txt).

Neighbor Discover messages allow network nodes to exchange information, as follows:

- Neighbor Solicitations allow a node to discover information about other nodes.
- Neighbor Advertisements are sent in response to Neighbor Solicitations. The Neighbor Advertisement contains information required by nodes to determine the link-layer address of the sender, and the sender's role on the network.
- IPv6 hosts use Router Solicitations to discover IPv6 routers. When a router receives a Router Solicitation, it responds immediately to the host.
- A router uses Router Advertisements to announce its presence on the network, and to provide its address prefix to neighbor devices. IPv6 hosts listen for Router Advertisements, and uses the information to build a list of default routers. Each host uses this information to perform auto-configuration of IPv6 addresses.
- Redirect messages are sent by IPv6 routers to inform hosts of a better first-hop address for a specific destination. Redirect messages are only sent by routers for unicast traffic, are only unicast to originating hosts, and are only processed by hosts.

Host versus router

Each IPv6 interface can be configured as a router node or a host node, as follows:

- A router node's IP address is configured manually. Router nodes can send Router Advertisements.
- A host node's IP address is auto-configured. Host nodes listen for Router Advertisements that convey information about devices on the network.

> **IP forwarding:** When IP forwarding is turned on. All IPv6 interfaces configured on the switch can forward packets.

You can configure each IPv6 interface as either a host node or a router node. You can manually assign an IPv6 address to an interface in host mode, or the interface can be assigned an IPv6 address by an upstream router, using information from router advertisements to perform stateless auto-configuration.

2.5.5 IPv6 support

IBM System Networking switches running IBM Networking OS V6.8 support IPv6 features included in the following IETF RFCs:

- RFC 2740 for OSPF
- RFCs 3306 and 3307 for dynamic IPv6 multicast addresses
- RFC 3810 for Multicast Listener Discovery (MLDv2)
- RFC 4301 for IPv6 security
- RFC 4302 for the IPv6 Authentication Header
- RFCs 2404, 2410, 2451, 3602, and 4303 for IPv6 Encapsulating Security Payload (ESP), including NULL encryption, CBC-mode 3DES and AES ciphers, and HMAC-SHA-1-96.
- RFCs 4306, 4307, 4718, and 4835 for Internet Key Exchange (IKEv2) and cryptography
- RFC 4552 for OSPFv3 IPv6 authentication
- RFC 5114 for Diffie-Hellman groups

To learn more about these RFCs, go to the following address and search for each RFC individually:

http://www.ietf.org/rfc/

2.6 Monitoring

In this section, we describe different features available in IBM System Networking switches that can be used to monitor characteristics of forwarded traffic or characteristics of the switch itself.

2.6.1 Port mirroring

You can use the IBM System Networking switches port mirroring feature to mirror (copy) the packets of a target port, and forward them to a monitoring port. Port mirroring functions for all Layer 2 and Layer 3 traffic on a port. This feature can be used as a troubleshooting tool or to enhance the security of your network.

For example, an intrusion detection system (IDS) server or other traffic sniffer device or analyzer can be connected to the monitoring port to detect intruders that attack the network.

IBM System Networking switches support a “many to one” mirroring model. As shown in Figure 2-11, selected traffic for ports 1 and 2 is being monitored by port 3. In the example, both ingress traffic and egress traffic on port 2 are copied and forwarded to the monitor. However, port 1 mirroring is configured so that only ingress traffic is copied and forwarded to the monitor. A device attached to port 3 can analyze the resulting mirrored traffic.

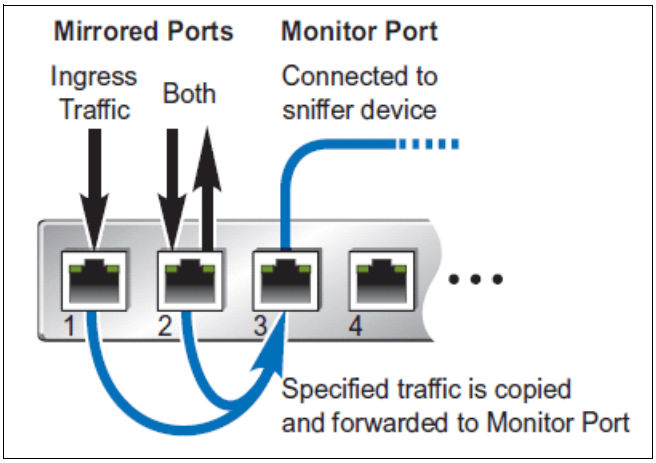

Figure 2-11 Mirroring ports

2.6.2 ACL-based mirroring

For regular ACLs (see 2.8.6, “Access control lists” on page 97) and VMaps (see 2.8.7, “VLAN maps” on page 100), packets that match an ACL on a specific port can be mirrored to another switch port for network diagnosis and monitoring.

The source port for the mirrored packets cannot be a portchannel, but may be a member of a portchannel.

The destination port to which packets are mirrored must be a physical port.

If the ACL or VMap has an action (permit, drop, and so on) assigned, it cannot be used to mirror packets for that ACL.

2.6.3 sFlow

IBM System Networking switches support sFlow technology for monitoring traffic in data networks. The switch includes an embedded sFlow agent that can be configured to provide continuous monitoring information of IPv4 traffic to a central sFlow analyzer.

The switch is responsible only for forwarding sFlow information. A separate sFlow analyzer is required elsewhere on the network to interpret sFlow data.

sFlow statistical counters

IBM System Networking switch can be configured to send network statistics to an sFlow analyzer at regular intervals. For each port, a polling interval of 5 - 60 seconds can be configured, or 0 (the default) can be set to disable this feature.

When polling is enabled, at the end of each configured polling interval, the switch reports general port statistics and port Ethernet statistics.

sFlow network sampling

In addition to statistical counters, IBM System Networking switches can be configured to collect periodic samples of the traffic data received on each port. For each sample, 128 bytes are copied, UDP-encapsulated, and sent to the configured sFlow analyzer.

For each port, the sFlow sampling rate can be configured to occur every 256 - 65536 packets, or set to 0 to disable (the default) this feature. A sampling rate of 256 means that one sample is taken for approximately every 256 packets received on the port. The sampling rate is statistical, however. It is possible to have more or fewer samples sent to the analyzer for any specific group of packets (especially under low traffic conditions). The actual sample rate becomes most accurate over time, and under higher traffic flow.

sFlow sampling has the following restrictions:

- Sample rate: The fastest sFlow sample rate is 1 out of every 256 packets.
- ACLs: sFlow sampling is performed before ACLs are processed. For ports configured both with sFlow sampling and one or more ACLs, sampling occurs regardless of the action of the ACL.
- Port mirroring: sFlow sampling does not occur on mirrored traffic. If sFlow sampling is enabled on a port that is configured as a port monitor, the mirrored traffic is not sampled.

sFlow sampling: Although sFlow sampling is not generally a processor -intensive operation, configuring fast sampling rates (such as once every 256 packets) on ports under heavy traffic loads can cause switch processor utilization to reach maximum. Use larger rate values for ports that experience heavy traffic.

2.6.4 Remote Monitoring (RMON)

Remote Monitoring (RMON) allows network devices to exchange network monitoring data. RMON allows the switch to perform the following functions:

- Track events and trigger alarms when a threshold is reached.
- Notify administrators by issuing a syslog message or SNMP trap.

RMON overview

The RMON MIB provides an interface between the RMON agent on the switch and an RMON management application. The RMON MIB is described in RFC 1757, found at:

http://www.ietf.org/rfc/rfc1757.txt

The RMON standard defines objects that are suitable for the management of Ethernet networks. The RMON agent continuously collects statistics and proactively monitors switch performance. You can use RMON to monitor traffic that flows through the switch.

2.7 High availability

Internet traffic consists of myriad services and applications that use the Internet Protocol (IP) for data delivery. However, IP is not optimized for all the various applications. High availability goes beyond IP and makes intelligent switching decisions to provide redundant network configurations.

2.7.1 Trunking

When using port trunk groups between two switches, as shown in Figure 2-12, you can create a virtual link between the switches, operating with combined throughput levels that depends on how many physical ports are included.

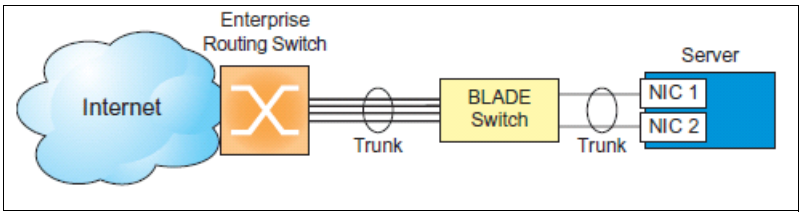

Figure 2-12 Trunking

Two trunk types are available:

- Static trunk groups (portchannel)
- Dynamic LACP trunk groups

Trunk groups are also useful for connecting a G8264 to third-party devices that support link aggregation, such as Cisco routers and switches with EtherChannel technology (not ISL trunking technology) and the Sun Quad Fast Ethernet Adapter. Trunk group technology is compatible with these devices when they are configured manually.

Trunk traffic is statistically distributed among the ports in a trunk group, based on various configurable options.

Also, because each trunk group is composed of multiple physical links, the trunk group is inherently fault tolerant. If one connection between the switches is available, the trunk remains active, and statistical load balancing is maintained whenever a port in a trunk group is lost or returned to service.

Configurable trunk hash algorithm

Traffic in a trunk group is statistically distributed among member ports by using a hash process where various address and attribute bits from each transmitted frame are recombined to specify the particular trunk port the frame uses.

The switch can be configured to use various hashing options. To achieve the most even traffic distribution, select options that exhibit a wide range of values for your particular network. Avoid hashing on information that is not present in the expected traffic, or which does not vary.

Link Aggregation Control Protocol

Link Aggregation Control Protocol (LACP) is an IEEE 802.3ad standard for grouping several physical ports into one logical port (known as a dynamic trunk group or Link Aggregation group) with any device that supports the standard. See the IEEE 802.3ad-2002 specification for a full description of the standard.

The 802.3ad standard allows standard Ethernet links to form a single Layer 2 link by using the Link Aggregation Control Protocol (LACP). Link aggregation is a method of grouping physical link segments of the same media type and speed in full duplex, and treating them as through they were part of a single, logical link segment. If a link in a LACP trunk group fails, traffic is reassigned dynamically to the remaining links of the dynamic trunk group.

A port's Link Aggregation Identifier (LAG ID) determines how the port can be aggregated. The LAG ID is constructed mainly from the system ID and the port's admin key, as follows:

- System ID: An integer value based on the switch's MAC address and the system priority assigned in the CLI.
- Admin key: A port's Admin key is an integer value (1 - 65535) that you can configure in the CLI. Each switch port that participates in the same LACP trunk group must have the same admin key value. The Admin key is local significant, which means the partner switch does not need to use the same Admin key value.

LACP automatically determines which member links can be aggregated and then aggregates them. It provides for the controlled addition and removal of physical links for the link aggregation. Up to 64 ports can be assigned to a single LAG, but only 16 ports can actively participate in the LAG at a time.

Each port on the switch can have one of the following LACP modes.

- Off (default)

  The user can configure this port in a regular static trunk group.

- Active

  The port can form an LACP trunk. This port sends LACPDU packets to partner system ports.

- Passive

  The port can form an LACP trunk. This port responds only to the LACPDU packets sent from an LACP active port.

Each active LACP port transmits LACP data units (LACPDUs), while each passive LACP port listens for LACPDUs. During LACP negotiation, the admin key is exchanged. The LACP trunk group is enabled if the information matches at both ends of the link. If the admin key value changes for a port at either end of the link, that port's association with the LACP trunk group is lost.

When the system is initialized, all ports by default are in LACP off mode and are assigned unique admin keys. To make a group of ports aggregatable, you assign them all the same admin key. You must set the port's LACP mode to active to activate LACP negotiation. You can set other port's LACP mode to passive, to reduce the amount of LACPDU traffic at the initial trunk-forming stage.

2.7.2 Virtual Link Aggregation Groups

In Figure 2-13, we show a typical data center design environment with access and aggregation layers.

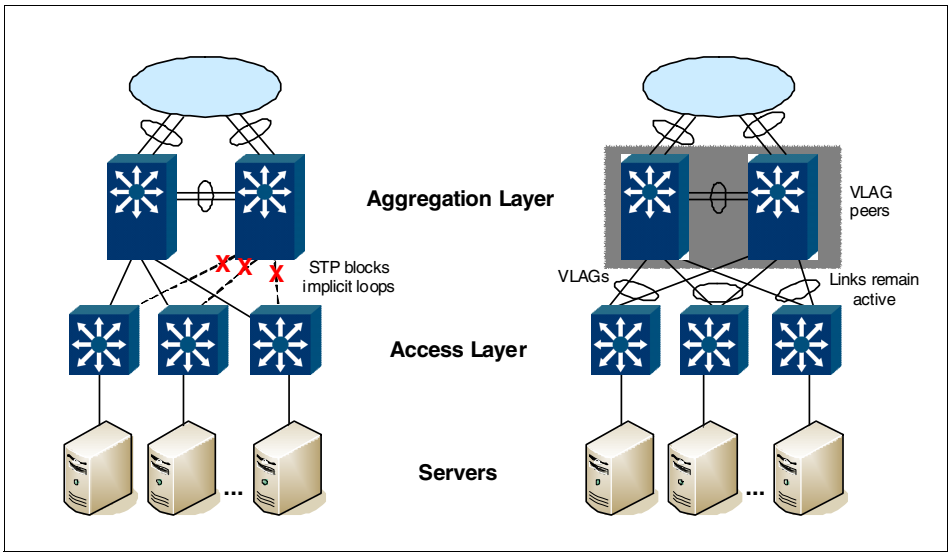

Figure 2-13 Spanning Tree Protocoal versus Virtual Link Aggregation Groups

As shown in Figure 2-13, a switch in the access layer may be connected to more than one switch in the aggregation layer to provide network redundancy. Typically, STP (see 2.1, “Virtual Local Area Networks” on page 52) is used to prevent forwarding loops by blocking redundant uplink paths. This configuration has the unwanted consequence of reducing the available bandwidth between the layers by as much as 50%. In addition, STP might be slow to resolve topology changes that occur during a link failure, and can result in considerable MAC address flooding. Using Virtual Link Aggregation Groups (VLAGs), the redundant uplinks remain active, using all available bandwidth.

Using the VLAG feature, switches can be paired as VLAG peers. The peer switches appear to the connecting device as a single virtual entity for establishing a multiport trunk. The VLAG-capable switches synchronize their logical view of the access layer port structure and internally prevent implicit loops. The VLAG topology also responds more quickly to link failure and does not result in unnecessary MAC flooding.

VLAGs are useful in multilayer environments for both uplink and downlink redundancy to any regular LAG-capable device. They can also be used in for active-active VRRP connections.

2.7.3 Hot links

For network topologies that require STP to be turned off, Hot Links provides basic link redundancy with fast recovery.

Hot Links consists of up to 25 triggers. A trigger consists of a pair of Layer 2 interfaces, each containing an individual port, trunk, or LACP adminkey. One interface is the Master, and the other is a Backup. Although the Master interface is set to the active state and forwards traffic, the Backup interface is set to the standby state and blocks traffic until the Master interface fails. If the Master interface fails, the Backup interface is set to active and forwards traffic. After the Master interface is restored, it changes to the standby state and blocks traffic until the Backup interface fails.

You may select a physical port, static trunk, or an LACP adminkey as a Hot Link interface.

Forward Delay

The Forward Delay timer allows Hot Links to monitor the Master and Backup interfaces for link stability before selecting one interface to change to the active state. Before the transition occurs, the interface must maintain a stable link for the duration of the Forward Delay interval.

For example, if you set the Forward Delay timer to 10 seconds, the switch selects an interface to become active only if a link remains stable for the duration of the Forward Delay period. If the link is unstable, the Forward Delay period starts again.

Preemption

You can configure the Master interface to resume the active state whenever it becomes available. With Hot Links preemption enabled, the Master interface transitions to the active state immediately upon recovery. The Backup interface immediately changes to the standby state. If Forward Delay is enabled, the transition occurs when an interface maintains link stability for the duration of the Forward Delay period.

FDB update

Use the FDB update option to notify other devices on the network about updates to the Forwarding Database (FDB). When you enable FDB update, the switch multicasts addresses in the FDB over the active interface, so that other devices on the network can learn the new path. The Hot Links FBD update option uses the station update rate to determine the rate at which to send FDB packets.

2.7.4 Fast Uplink Convergence

Fast Uplink Convergence enables IBM System Networking switches to quickly recover from the failure of the primary link or trunk group in a Layer 2 network using STP/PVST+ mode. Normal recovery can take up to 50 seconds, while the backup link transitions from Spanning Tree Blocking to Listening to Learning and then Forwarding states. With Fast Uplink Convergence enabled, the IBM System Networking switch immediately places the secondary path into Forwarding state, and multicasts the addresses in the FDB and ARP table over the secondary link so that upstream switches can learn the new path.

2.7.5 NIC teaming and Layer 2 failover

The primary application for Layer 2 Failover is to support Network Adapter Teaming. With Network Adapter Teaming, all the NICs on each server share an IP address, and are configured into a team. One NIC is the primary link, and the other is a standby link. For more details, see the documentation for your Ethernet adapter.

Failover: Only two links per server can be used for Layer 2 Trunk Failover (one primary and one backup). Network Adapter Teaming allows only one backup NIC for each server blade.

Monitoring trunk links

Layer 2 Failover can be enabled on any trunk group in IBM System Networking switches, including LACP trunks. Trunks can be added to failover trigger groups. Then, if some specified number of monitor links fail, the switch disables all the control ports in the switch. When the control ports are disabled, it causes the NIC team on the affected servers to fail over from the primary to the backup NIC. This process is called a failover event.

When the appropriate number of links in a monitor group return to service, the switch enables the control ports. This configuration causes the NIC team on the affected servers to fail back to the primary switch (unless Auto-Fallback is disabled on the NIC team). The backup switch processes traffic until the primary switch's control links come up, which can take up to 5 seconds.

Figure 2-14 is a simple example of Layer 2 Failover. One switch is the primary, and the other is used as a backup. In this example, all ports on the primary switch belong to a single trunk group, with Layer 2 Failover enabled, and Failover Limit set to 2. If two or fewer links in trigger 1 remain active, the switch temporarily disables all control ports. This action causes a failover event on Server 1 and Server 2.

This feature is also referred to as *Uplink Failure Detection*. The switch constantly monitors the port or port trunk group to the Core Network. When a failure is detected, the switch disables the pre-configured ports connected to the servers.

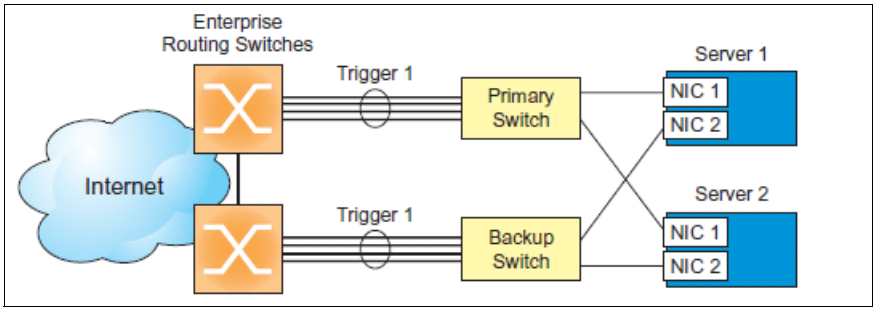

Figure 2-14 Basic Layer 2 Failover

Setting the failover limit

You can use the failover limit to specify the minimum number of operational links required within each trigger before the trigger initiates a failover event. For example, if the limit is two, a failover event occurs when the number of operational links in the trigger is two or fewer. When you set the limit to zero, the switch triggers a failover event only when no links in the trigger are operational.

Manually monitoring port links

You can use the Manual Monitor to configure a set of ports or trunks to monitor for link failures (a monitor list), and another set of ports or trunks to disable when the trigger limit is reached (a control list). When the switch detects a link failure on the monitor list, it automatically disables the items in control list. When server ports are disabled, the corresponding server's network adapter can detect the disabled link, and trigger a network-adapter failover to another port or trunk on the switch, or another switch.

The switch automatically enables the control list items when the monitor list items return to service.

Layer 2 Failover with other features

Layer 2 Failover works together with static trunks, LACP, and STP.

Static trunks

When you add a portchannel (static trunk group) to a failover trigger, any ports in that trunk become members of the trigger. You can add up to 64 static trunks to a failover trigger, using manual monitoring.

Link Aggregation Control Protocol

LACP allows the switch to form dynamic trunks. You can use the admin key to add up to two LACP trunks to a failover trigger by using automatic monitoring. When you add an admin key to a trigger, any LACP trunk with that admin key becomes a member of the trigger.

Spanning Tree Protocol

If STP is enabled on the ports in a failover trigger, the switch monitors the port STP state rather than the link state. A port failure results when STP is not in a Forwarding state (such as Listening, Learning, Blocking, or No Link). The switch automatically disables the appropriate control ports.

When the switch determines that ports in the trigger are in the STP Forwarding state, then it automatically enables the appropriate control ports. The switch fails back to normal operation.

2.7.6 Virtual Router Redundancy Protocol

IBM System Networking switches support IPv4 high-availability network topologies through an enhanced implementation of the Virtual Router Redundancy Protocol (VRRP).

Virtual Router Redundancy Protocol overview

In a high-availability network topology, no device can create a single point of failure for the network or force a single point-of-failure to any other part of the network. This situation means that your network remains in service despite the failure of any single device. To achieve this goal usually requires redundancy for all vital network components.

VRRP enables redundant router configurations within a LAN, providing alternative router paths for a host to eliminate single points of failure within a network. Each participating VRRP-capable routing device is configured with the same virtual router IPv4 address and ID number. One of the virtual routers is elected as the master, based on a number of priority criteria, and assumes control of the shared virtual router IPv4 address. If the master fails, one of the backup virtual routers takes control of the virtual router IPv4 address and actively processes traffic addressed to it.

With VRRP, Virtual Interface Routers (VIR) allow two VRRP routers to share an IP interface across the routers. VIRs provide a single Destination IPv4 (DIP) address for upstream routers to reach various servers, and provide a virtual default gateway for the servers.

Virtual Router Redundancy Protocol components

Each physical router that runs VRRP is known as a *VRRP router*.

Virtual router

Two or more VRRP routers can be configured to form a virtual router (For more details, see RFC 2338 at http://www.ietf.org/rfc/rfc2338.txt). Each VRRP router may participate in one or more virtual routers. Each virtual router consists of a user-configured virtual router identifier (VRID) and an IPv4 address.

Virtual router MAC address

The VRID is used to build the virtual router MAC Address. The five highest-order octets of the virtual router MAC Address are the standard MAC prefix (00-00-5E-00-01) defined in RFC 2338. The VRID is used to form the lowest-order octet.

Owners and renters

Only one of the VRRP routers in a virtual router may be configured as the IPv4 address owner. This router has the virtual router's IPv4 address as its real interface address. This router responds to packets addressed to the virtual router's IPv4 address for ICMP pings, TCP connections, and so on.

There is no requirement for any VRRP router to be the IPv4 address owner. Most VRRP installations choose not to implement an IPv4 address owner. For the purposes of this chapter, VRRP routers that are not the IPv4 address owner are called *renters*.

Master and backup virtual router

Within each virtual router, one VRRP router is selected to be the virtual router master. For an explanation of the selection process, see "Selecting the master VRRP router" on page 83.

Virtual router master: If the virtual IPv4 address owner is available, it always becomes the virtual router master.

The virtual router master forwards packets sent to the virtual router. It also responds to Address Resolution Protocol (ARP) requests sent to the virtual router's IPv4 address. Finally, the virtual router master sends out periodic advertisements to inform other VRRP routers that it is alive and what its priority is.

Within a virtual router, the VRRP routers not selected to be the master are known as *virtual router backups*. Should the virtual router master fail, one of the virtual router backups becomes the master, and assumes its responsibilities.

Virtual Interface Router

At Layer 3, a VIR allows two VRRP routers to share an IP interface across the routers. VIRs provide a single DIP address for upstream routers to reach various destination networks, and provide a virtual default gateway.

VIR considerations: Every VIR must be assigned to an IP interface, and every IP interface must be assigned to a VLAN. If no port in a VLAN has a link up, the IP interface of that VLAN is down, and if the IP interface of a VIR is down, that VIR goes into INIT state.

Virtual Router Redundancy Protocol operation

Only the virtual router master responds to ARP requests. Therefore, the upstream routers only forward packets destined to the master. The master also responds to ICMP **`ping`** requests. The backup does not forward any traffic, nor does it respond to ARP requests.

If the master is not available, the backup becomes the master and takes over responsibility for packet forwarding and responding to ARP requests.

Selecting the master VRRP router

Each VRRP router is configured with a priority 1 - 254. A bidding process determines which VRRP router is or becomes the master, that is, the VRRP router with the highest priority.

The master periodically sends advertisements to an IPv4 Multicast address. IF the backups receive these advertisements, they remain in the backup state. If a backup does not receive an advertisement for three advertisement intervals, it initiates a bidding process to determine which VRRP router has the highest priority and takes over as master.

If, at any time, a backup determines that it has higher priority than the current master does, it can preempt the master and become the master itself, unless configured not to do so. In preemption, the backup assumes the role of master and begins to send its own advertisements. The current master sees that the backup has higher priority and stops functioning as the master.

A backup router can stop receiving advertisements for one of two reasons: the master can be down, or all communications links between the master and the backup can be down. If the master fails, it is clearly desirable for the backup (or one of the backups, if there is more than one) to become the master.

Two masters: If the master is healthy but communication between the master and the backup fails, there are then two masters within the virtual router. To prevent this situation from happening, configure redundant links to be used between the switches that form a virtual router.

Failover methods

With service availability on the Internet becoming a major concern, service providers are increasingly deploying Internet traffic control devices, such as application switches, in redundant configurations. IBM System Networking high availability configurations are based on VRRP. The IBM System Networking implementation of VRRP includes proprietary extensions.

Active-active redundancy

In an active-active configuration, shown Figure 2-15, two switches provide redundancy for each other, with both active at the same time. Each switch processes traffic on a different subnet. When a failure occurs, the remaining switch can process traffic on all subnets.

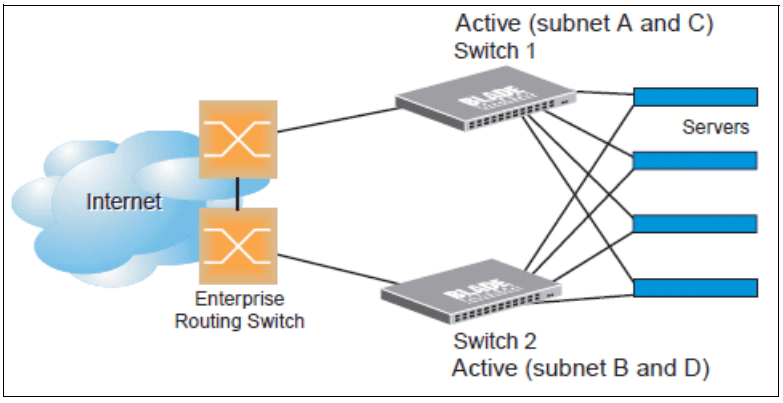

Figure 2-15 Active-active redundancy

Virtual router group

The virtual router group ties all virtual routers on the switch together as a single entity. As members of a group, all virtual routers on the switch (and therefore the switch itself) are in either a master or standby state.

A virtual router group has the following characteristics:

- When enabled, all virtual routers behave as one entity, and all group settings override any individual virtual router settings.
- All individual virtual routers, after the VRRP group is enabled, assume the group's tracking and priority.
- When one member of a VRRP group fails, the priority of the group decreases, and the state of the entire switch changes from Master to Standby.

Each VRRP advertisement can include up to 16 addresses. All virtual routers are advertised within the same packet, conserving processing and buffering resources.

High-availability configurations

Figure 2-16 shows an example configuration where two IBM System Networking switches are used as VRRP routers in an active-active configuration. In this configuration, both switches respond to packets.

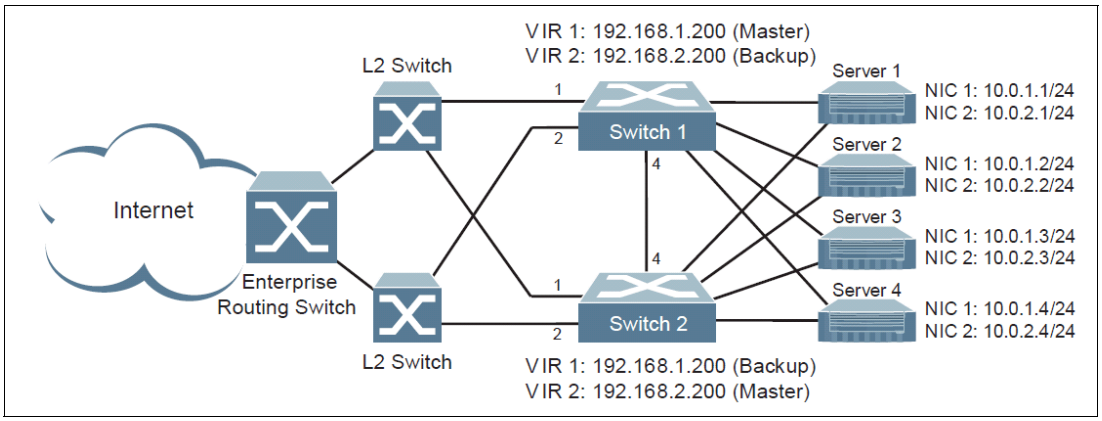

Figure 2-16 Active-active configuration with VRRP

Although this example shows only two switches, there is no limit of the number of switches used in a redundant configuration. It is possible to implement an active-active configuration across all the VRRP-capable switches in a LAN.

Each VRRP-capable switch in an active-active configuration is autonomous. Switches in a virtual router do not need to be identically configured.

In the scenario illustrated in Figure 2-16, traffic destined for IPv4 address 10.0.1.1 is forwarded through the Layer 2 switch at the top of the drawing, and ingresses Switch 1 on port 1. Return traffic uses default gateway 1 (192.168.1.1).

If the link between Switch 1 and the Layer 2 switch fails, Switch 2 becomes the master because it has a higher priority. Traffic is forwarded to Switch 2, which forwards it to Switch 1 through port 4.

Return traffic uses default gateway 2 (192.168.2.1), and is forwarded through the Layer 2 switch at the bottom of the drawing.

VRRP high availability with VLAGs

VRRP can be used in conjunction with VLAGs and LACP-capable servers and switches to provide seamless redundancy (Figure 2-17).

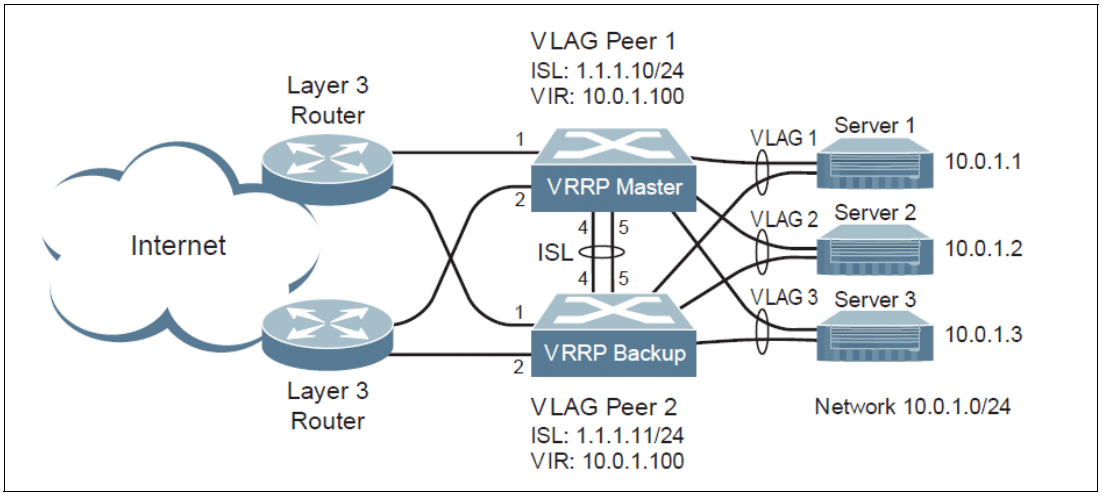

Figure 2-17 Active-active configuration using VRRP and VLAGs

2.7.7 Active Multipath Protocol

You can use Active MultiPath Protocol (AMP) to connect three switches in a loop topology, and load-balance traffic across all uplinks (no blocking). When an AMP link fails, upstream communication continues over the remaining AMP link. After the failed AMP link re-establishes connectivity, communication resumes to its original flow pattern.

AMP is supported over Layer 2 only. Layer 3 routing is not supported. STP is not required in an AMP Layer 2 domain. STP BPDUs are not forwarded over the AMP links, and any BPDU packets received on AMP links are dropped.

Each AMP group contains two aggregator switches and one access switch. Aggregator switches support up to 22 AMP groups. Access switches support only one AMP group.

Figure 2-18 shows a typical AMP topology, with two aggregators that support a number of AMP groups.

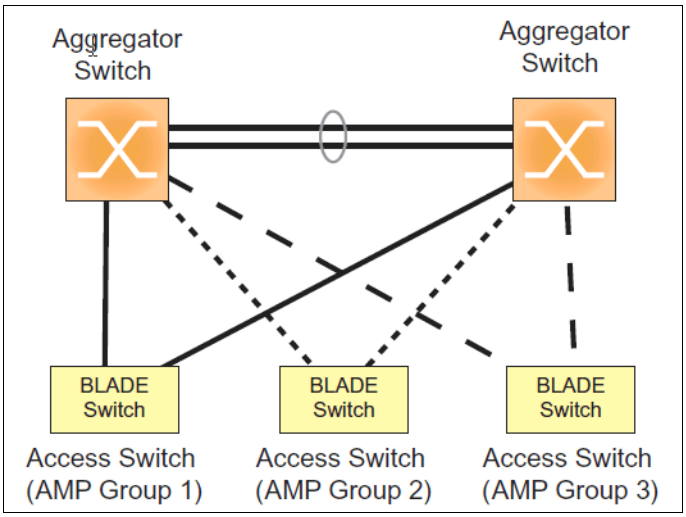

Figure 2-18 Active Multipath Protocol topology

Each AMP group requires two links on each switch. Each AMP link consists of a single port, a static trunk group, or an LACP trunk group. Local non-AMP ports can communicate through local Layer 2 switching without passing traffic through the AMP links. No two switches in the AMP loop can have another active connection between them through a non-AMP switch.

Each AMP switch has a priority value (1 - 255). The switch with the lowest priority value has the highest precedence over the other switches. If there is a conflict between switch priorities, the switch with lowest MAC address has the highest precedence.

AMP and access switches: For correct AMP operation, all access switches should be configured with a higher priority value (lower precedence) than the aggregators. Otherwise, some AMP control packets may be sent to access switches, even when their AMP groups are disabled.

When the AMP loop is broken, the STP port states are set to forwarding or blocking, depending on the switch priority and port/trunk precedence, as follows:

- An aggregator's port/trunk has higher precedence over an access switch's port/trunk.
- Static trunks have highest precedence, followed by LACP trunks, then physical ports.
- Between two static trunks, the trunk with the lower trunk ID has higher precedence.
- Between two LACP trunks, the trunk with the lower admin key has higher precedence.
- Between two ports, the port with the lowest port number has higher precedence.

Health checks

An AMP keepalive message is passed periodically from each switch to its neighbors in the AMP group. The keepalive message is a BPDU-like packet that passes on an AMP link even when the link is blocked by STP. The keepalive message carries status information about AMP ports/trunks, and is used to verify that a physical loop exists.

An AMP link is considered healthy if the switch receives an AMP keepalive message on that link. An AMP link is considered unhealthy if a number of consecutive AMP keepalive messages have not been received recently on that link.

FDB flush

When an AMP port/trunk is the blocking state, FDB flush is performed on that port/trunk. Any time there is a change in the data path for an AMP group, the FDB entries associated with the ports in the AMP group are flushed. This situation ensures that communication is not blocked while obsolete FDB entries are aged out.

FDB flush is performed when an AMP link goes down, and when an AMP link comes up.

2.7.8 Stacking

A *stack* is a group of up to eight IBM System Networking switches (the stacking is supported only on Virtual Fabric 10Gb Switch Module devices) that work together as a unified system. Because the multiple members of a stack acts as a single switch entity with distributed resources, high-availability topologies can be more easily achieved.

In Figure 2-19, a simple stack using two switches provides full redundancy in the event that either switch fails. As shown with the servers in the example, stacking permits ports within different physical switches to be trunked together, further enhancing switch redundancy.

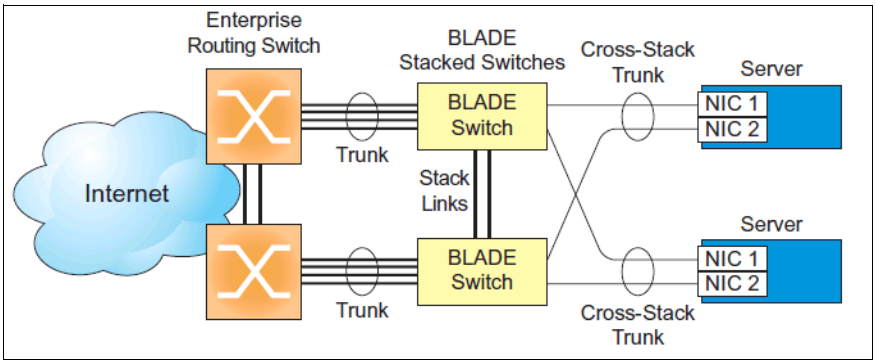

Figure 2-19 High-availability topology using stacking

A stack has the following properties, regardless of the number of switches included:

- The network views the stack as a single entity.
- The stack can be accessed and managed as a whole using standard switch IP interfaces configured with IPv4 addresses.
- After the stacking links are established, the number of ports available in a stack equals the total number of remaining ports of all the switches that are part of the stack.
- The number of available IP interfaces, VLANs, trunks, trunk links, and other switch attributes are not aggregated among the switches in a stack. The totals for the stack as a whole are the same as for any single switch configured in stand-alone mode.

Stacking requirements

Before IBM System Networking switches can form a stack, they must meet the following requirements:

- All switches must be the same model (Virtual Fabric 10Gb Switch Module).
- Each switch must be installed with IBM Networking OS Version 6.5 or later. The same release version is not required, as the master switch pushes a firmware image to each switch that is part of the stack.

- The preferred stacking topology is a bidirectional ring (Figure 2-20). To achieve this topology, two external 10Gb Ethernet ports on each switch must be reserved for stacking. By default, the first two 10Gb Ethernet ports are used.

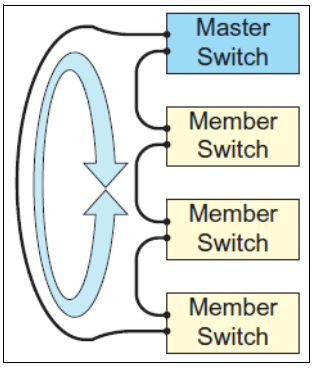

Figure 2-20 Stacking connection

- The cables used for connecting the switches in a stack carry low-level, inter-switch communications and cross-stack data traffic critical to shared switching functions. Always maintain the stability of stack links in order to avoid internal stack reconfiguration.

2.8 Security

This section presents various security features and protocols available on IBM System Networking switches.

2.8.1 Private VLANs

Private VLANs (see 2.1, “Virtual Local Area Networks” on page 52) provide Layer 2 isolation between the ports within the same broadcast domain. Private VLANs can control traffic within a VLAN domain, and provide port-based security for host servers.

Use Private VLANs to partition a VLAN domain into subdomains. Each subdomain is composed of one primary VLAN and one secondary VLAN, as follows:

- Primary VLAN: Carries unidirectional traffic downstream from promiscuous ports. Each Private VLAN has only one primary VLAN. All ports in the Private VLAN are members of the primary VLAN.
- Secondary VLAN: Secondary VLANs are internal to a private VLAN domain, and are defined as follows:
  - Isolated VLAN: Carries unidirectional traffic upstream from the host servers toward ports in the primary VLAN and the gateway. Each Private VLAN can contain only one Isolated VLAN.
  - Community VLAN: Carries upstream traffic from ports in the community VLAN to other ports in the same community, and to ports in the primary VLAN and the gateway. Each. Private VLAN can contain multiple community VLANs.

After you define the primary VLAN and one or more secondary VLANs, you map the secondary VLANs to the primary VLAN.

Private VLAN ports

Private VLAN ports are defined as follows:

- Promiscuous: A promiscuous port is an external port that belongs to the primary VLAN. The promiscuous port can communicate with all the interfaces, including ports in the secondary VLANs (Isolated VLAN and Community VLANs). Each promiscuous port can belong to only one Private VLAN.
- Isolated: An isolated port is a host port that belongs to an isolated VLAN. Each isolated port has complete Layer 2 separation from other ports within the same private VLAN (including other isolated ports), except for the promiscuous ports.
  - Traffic sent to an isolated port is blocked by the Private VLAN, except the traffic from promiscuous ports.
  - Traffic received from an isolated port is forwarded only to promiscuous ports.
- Community: A community port is a host port that belongs to a community VLAN. Community ports can communicate with other ports in the same community VLAN, and with promiscuous ports. These interfaces are isolated at Layer 2 from all other interfaces in other communities and from isolated ports within the Private VLAN.

Only external ports are promiscuous ports. Only internal ports may be isolated or community ports.

2.8.2 Securing administration

In this section, we present the different features and protocols used to secure administrative access.

Secure Shell and Secure Copy

Because using Telnet does not provide a secure connection for managing an IBM System Networking switch, Secure Shell (SSH) and Secure Copy (SCP) features are included for IBM System Networking switch management. SSH and SCP use secure tunnels to encrypt and secure messages between a remote administrator and the switch.

SSH is a protocol that enables remote administrators to log on securely to the switch over a network to execute management commands.

SCP is typically used to copy files securely from one machine to another. SCP uses SSH for encryption of data on the network. On a switch, SCP is used to download and upload the switch configuration through secure channels.

Although SSH and SCP are disabled by default, enabling and using these features provides the following benefits:

- Identifying the administrator using a user name and password
- Authentication of remote administrators
- Authorization of remote administrators
- Determining the permitted actions and customizing service for individual administrators
- Encryption of management messages
- Encrypting messages between the remote administrator and switch
- Secure copy support

User access control

IBM System Networking switch allows an administrator to define user accounts that permit users to perform operation tasks through the switch CLI commands. After user accounts are configured and enabled, the switch requires user name and password authentication.

For example, an administrator can assign a user, who can then log on to the switch and perform operational commands (effective only until the next switch reboot).

Considerations for configuring user accounts

Consider the following items when configuring use accounts:

- A maximum of 10 user IDs are supported on the switch.
- IBM System Networking switch supports user support for Console, Telnet, BBI, and SSHv1/v2 access to the switch.
- If RADIUS authentication is used, the user password on the Radius server overrides the user password on the switch. The password change command modifies only the user switch password on the switch and has no effect on the user password on the Radius server. Radius authentication and user password cannot be used concurrently to access the switch.
- Passwords can be up to 128 characters in length for TACACS, RADIUS, Telnet, SSH, Console, and web access.

Protected Mode

Protected Mode settings (available only for Virtual Fabric 10Gb Switch Module for IBM BladeCenter) allow the switch administrator to block the management module from making configuration changes that affect switch operation. The switch retains control over those functions.

The following management module functions are disabled when Protected Mode is turned on:

- External Ports: Enabled/Disabled
- External management over all ports: Enabled/Disabled
- Restore Factory Defaults
- New Static IP Configuration

In IBM Networking OS V6.5, the configuration of these functions is restricted to the local switch when you turn Protected Mode on. With new releases, the number of functions over which you have an individual control is increasing.

Protected mode: Before you turn Protected Mode on, make sure that external management (Telnet) access to one of the switch's IP interfaces is enabled.

2.8.3 Authentication and authorization protocols

In this section, we provide information about the two most common authentication and authorization protocols (Radius and TACACS+) and support for those protocols on IBM System Networking switches.

RADIUS authentication and authorization

IBM System Networking switch supports the RADIUS (Remote Authentication Dial-in User Service) method to authenticate and authorize remote administrators for managing the switch. This method is based on a client/server model. The Remote Access Server (RAS), the switch, is a client to the back-end database server. A remote user (the remote administrator) interacts only with the RAS, not the back-end server and database.

RADIUS authentication consists of the following components:

- A protocol with a frame format that uses UDP over IP (based on RFC 2138, found at http://www.ietf.org/rfc/rfc2138.txt and RFC 2866, found at http://www.ietf.org/rfc/rfc2866.txt)
- A centralized server that stores all the user authorization information
- A client; in this case, the switch

The IBM System Networking switch, acting as the RADIUS client, communicates with the RADIUS server to authenticate and authorize a remote administrator by using the protocol definitions specified in RFC 2138 and RFC 2866. Transactions between the client and the RADIUS server are authenticated by using a shared key that is not sent over the network. In addition, the remote administrator passwords are sent encrypted between the RADIUS client (the switch) and the back-end RADIUS server.

How RADIUS authentication works

RADIUS authentication uses the following steps:

1. A remote administrator connects to the switch and provides a user name and password.
2. Using the Authentication/Authorization protocol, the switch sends a request to the authentication server.
3. The authentication server checks the request against the user ID database.
4. Using the RADIUS protocol, the authentication server instructs the switch to grant or deny administrative access.

RADIUS authentication features in IBM System Networking switches

IBM System Networking switches support the following RADIUS authentication features:

- Supports RADIUS client on the switch, based on the protocol defined in RFC 2138 and RFC 2866.
- Allows a RADIUS secret password that is up to 32 bytes and less than 16 octets.
- Supports a secondary authentication server so that when the primary authentication server is unreachable, the switch can send client authentication requests to the secondary authentication server.
- Supports user-configurable RADIUS server retry and timeout values:
  - Timeout value: 1 - 10 seconds
  - Retries: 1 - 3

  The switch times out if it does not receive a response from the RADIUS server after 1 - 3 attempts. The switch also automatically tries to connect to the RADIUS server before it declares the server down.
- Supports a user-configurable RADIUS application port.

  The default is 1812/UDP-based, as described in RFC 2138, found at http://www.ietf.org/rfc/rfc2138.txt. Port 1645 is also supported.

- Supports a user-configurable RADIUS application port. The default is UDP port 1645. UDP port 1812, based on RFC 2138, is also supported.
- Allows network administrator to define privileges for one or more specific users to access the switch at the RADIUS user database.

Switch user accounts

The user accounts listed in Table 2-1 can be defined in the RADIUS server dictionary file.

Table 2-1 User access levels

User account	Description and tasks performed	Password
User	The user has no direct responsibility for switch management. The user can view all switch status information and statistics, but cannot make any configuration changes to the switch.	user
Operator	The operator manages all functions of the switch. The operator can reset ports, except the management port.	oper
Administrator	The super-user administrator has complete access to all commands, information, and configuration commands on the switch, including the ability to change both the user and administrator passwords.	admin

RADIUS attributes for IBM Networking OS user privileges

When the user logs in, the switch authenticates the user's level of access by sending the RADIUS access request, that is, the client authentication request, to the RADIUS authentication server.

If the remote user is successfully authenticated by the authentication server, the switch verifies the *privileges* of the remote user and authorizes the appropriate access. The administrator can allow secure *back door* access through Telnet/SSH/BBI (or Telnet, SSH, HTTP, and HTTPS in the case of Virtual Fabric 10Gb Switch Module for IBM BladeCenter). Secure back door provides switch access when the RADIUS servers cannot be reached. You always can access the switch through the console port, by using the noradius user ID and the administrator password, whether the secure back door is enabled or not.

All user privileges, other than the ones assigned to the administrator, must be defined in the RADIUS dictionary. RADIUS attribute 6, which is built into all RADIUS servers, defines the administrator. The file name of the dictionary is RADIUS vendor-dependent. The following RADIUS attributes are defined for IBM Networking OS user privileges levels:

Table 2-2 IBM System Networking switches proprietary attributes for RADIUS

User name/access	User-Service-Type	Value
User	Vendor-supplied	255
Operator	Vendor-supplied	252
Admin	Vendor-supplied	6

TACACS+ authentication

IBM Networking OS supports authentication, authorization, and accounting with networks using the Cisco Systems TACACS+ protocol. The switch functions as the Network Access Server (NAS) by interacting with the remote client and initiating authentication and authorization sessions with the TACACS+ access server. The remote user is defined as someone that requires management access to the VFSM either through a data or management port.

TACACS+ offers the following advantages over RADIUS:

- TACACS+ uses TCP-based connection-oriented transport, where RADIUS is UDP-based. TCP offers a connection-oriented transport, where UDP offers best-effort delivery. RADIUS requires additional programmable variables, such as retransmit attempts and timeouts to compensate for best-effort transport, but it lacks the level of built-in support that a TCP transport offers.
- TACACS+ offers full packet encryption where RADIUS offers password-only encryption in authentication requests.
- TACACS+ separates authentication, authorization, and accounting.

How TACACS+ authentication works

TACACS+ works similar to RADIUS authentication:

1. A remote administrator connects to the switch and provides a user name and password.
2. Using the Authentication/Authorization protocol, the switch sends request to the authentication server.
3. The authentication server checks the request against the user ID database.
4. Using the TACACS+ protocol, the authentication server instructs the switch to grant or deny administrative access.

During a session, if additional authorization checking is needed, the switch checks with a TACACS+ server to determine if the user is granted permission to use a particular command.

TACACS+ authentication features in IBM System Networking switches

Authentication is the action of determining the identity of a user, and is generally done when the user first attempts to log on to a device or gain access to its services. IBM System Networking switches support ASCII inbound login to the device. PAP, CHAP, and ARAP login methods, TACACS+ change password requests, and one-time password authentication are not supported.

Authorization

Authorization is the action of determining a user's privileges on the device, and usually takes place after authentication.

The default mapping between TACACS+ authorization levels and IBM Networking OS management access levels is shown in Table 2-3.

Table 2-3 Default TACACS+ authorization levels

User access level	TACACS+ level
user	0
oper	3
admin	6

If the remote user is successfully authenticated by the authentication server, the switch verifies the privileges of the remote user and authorizes the appropriate access. The administrator may allow secure back door access through Telnet/SSH. Secure back door provides switch access when the TACACS+ servers cannot be reached.

Accounting

Accounting is the action of recording a user's activities on the device for the purposes of billing and security. It follows the authentication and authorization actions. If the authentication and authorization is not performed through TACACS+, there are no TACACS+ accounting messages sent out.

You can use TACACS+ to record and track software login access, configuration changes, and interactive commands.

LDAP authentication and authorization

IBM System Networking switches support the Lightweight Directory Access Protocol (LDAP) method to authenticate and authorize remote administrators to manage the switch. LDAP is based on a client/server model.

The switch acts as a client to the LDAP server. A remote user (the remote administrator) interacts only with the switch, not the back-end server and database.

LDAP authentication consists of the following components:

- A protocol with a frame format that uses TCP over IP
- A centralized server that stores all the user authorization information
- A client, in this case, the switch

Each entry in the LDAP server is referenced by its Distinguished Name (DN). The DN consists of the user-account name concatenated with the LDAP domain name. If the user-account name is John, the following is an example DN:

```
uid=John,ou=people,dc=domain,dc=com
```

2.8.4 MAC address notification

MAC address notification is a feature that causes a switch to generate a syslog message when a MAC address is added or removed from the MAC address table. This feature is useful for tracking hosts as they change the ports they are connected to.

2.8.5 802.1x Port-based network access control

Port-based network access control provides a means of authenticating and authorizing devices attached to a LAN port that has point-to-point connection characteristics. It prevents access to ports that fail authentication and authorization. This feature provides security to ports of IBM System Networking Switch Module that connect to blade servers.

Extensible Authentication Protocol over LAN

IBM Networking OS can provide user-level security for its ports by using the IEEE 802.1X protocol, which is a more secure alternative to other methods of port-based network access control. Any device attached to an 802.1X-enabled port that fails authentication is prevented access to the network and denied services offered through that port.

The 802.1X standard describes port-based network access control by using Extensible Authentication Protocol over LAN (EAPoL). EAPoL provides a means of authenticating and authorizing devices attached to a LAN port that has point-to-point connection characteristics and of preventing access to that port in cases of authentication and authorization failures.

EAPoL is a client-server protocol that has the following components:

- Supplicant or Client

  The Supplicant is a device that requests network access and provides the required credentials (user name and password) to the Authenticator and the Authenticator Server.

- Authenticator

  The Authenticator enforces authentication and controls access to the network. The Authenticator grants network access based on the information provided by the Supplicant and the response from the Authentication Server. The Authenticator acts as an intermediary between the Supplicant and the Authentication Server: requesting identity information from the client, forwarding that information to the Authentication Server for validation, relaying the server's responses to the client, and authorizing network access based on the results of the authentication exchange. The IBM System Networking switch acts as an Authenticator.

- Authentication Server

  The Authentication Server validates the credentials provided by the Supplicant to determine whether the Authenticator should grant access to the network. The Authentication Server may be co-located with the Authenticator. The VFSM relies on external RADIUS servers for authentication.

Upon a successful authentication of the client by the server, the 802.1X-controlled port transitions from unauthorized to authorized state, and the client is allowed full access to services through the port. When the client sends an EAP-Logoff message to the authenticator, the port transitions from an authorized to unauthorized state.

EAPoL authentication process

The clients and authenticators communicate by using Extensible Authentication Protocol (EAP), which was originally designed to run over PPP, and for which the IEEE 802.1X Standard defined an encapsulation method over Ethernet frames, called EAP over LAN (EAPoL).

Figure 2-21 shows a typical message exchange initiated by the client.

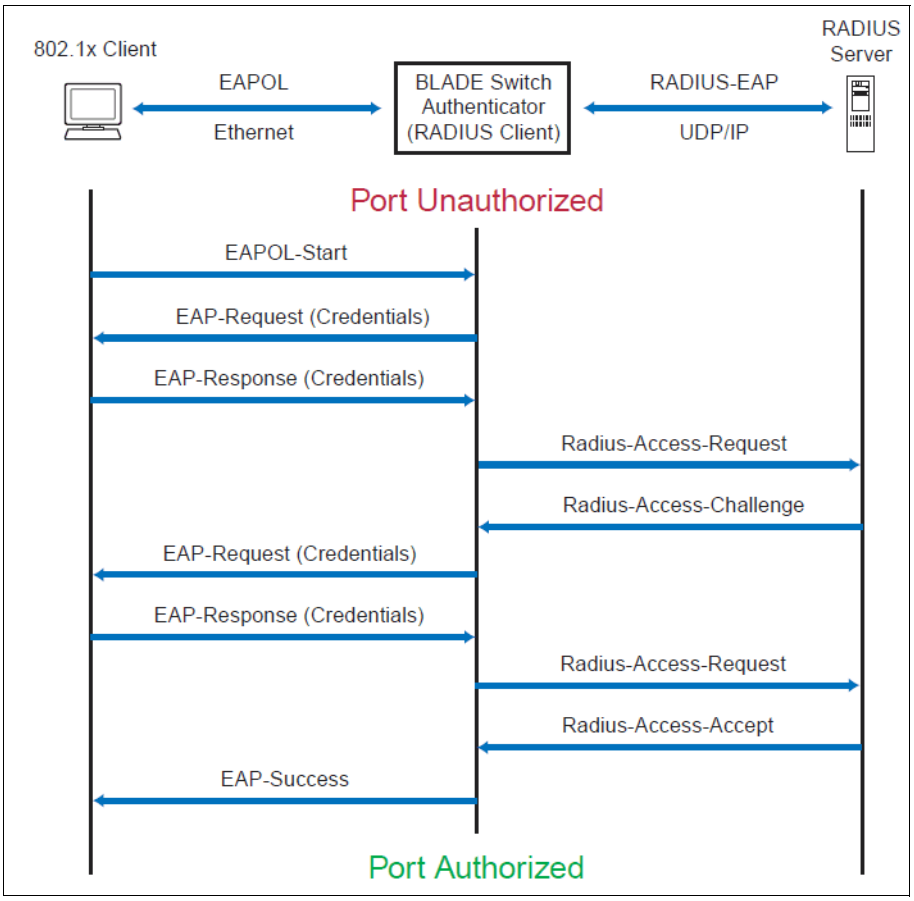

Figure 2-21 Authenticating a port by using EAPoL

2.8.6 Access control lists

Access control lists (ACLs) are filters that permit or deny traffic for security purposes. They can also be used with QoS to classify and segment traffic to provide different levels of service to different traffic types. Each filter defines the conditions that must match for inclusion in the filter, and also the actions that are performed when a match is made.

IBM System Networking switches running IBM Networking OS V6.8 support the following ACLs:

- Regular ACLs:

  Up to 256 ACLs are supported for networks that use IPv4 addressing.
- IPv6 ACLs:

  Up to 128 ACLs are supported for networks that use IPv6 addressing.
- VLAN Maps (VMaps):

  Up to 128 VLAN Maps are supported for attaching filters to VLANs rather than ports.

Summary of packet classifiers

You can use ACLs to classify packets according to various content in the packet header (such as the source address, destination address, source port number, destination port number, and others). Once classified, packet flows can be identified for more processing.

You can use regular ACLs, IPv6 ACLs, and VMaps to classify packets based on the following packet attributes:

- Ethernet header options (for regular ACLs and VMaps only)
  - Source MAC address
  - Destination MAC address
  - VLAN number and mask
  - Ethernet type (ARP, IP, IPv6, MPLS, RARP, and so on)
  - Ethernet priority (the IEEE 802.1p priority)
- IPv4 header options (for regular ACLs and VMaps only)
  - Source IPv4 address and subnet mask
  - Destination IPv4 address and subnet mask
  - Type of Service value
  - IP protocol number or name, as shown in Table 2-4

Table 2-4 Well-known protocol types

Number	Protocol name
1	ICMP
2	IGMP
6	TCP
17	UDP
89	OSPF
112	VRRP

- IPv6 header options (for IPv6 ACLs only)
  - Source IPv6 address and prefix length
  - Destination IPv6 address and prefix length
  - Next Header value
  - Flow Label value
  - Traffic Class value
- TCP/UDP header options (for all ACLs)
  - TCP/UDP application source port and mask, as shown in Table 2-5
  - TCP/UDP application destination port, as shown in Table 2-5

Table 2-5 Well-known application ports

Port	Application	Port	Application	Port	Application
20/udp	ftp-data	79	finger	179	bgp
21	FTP	80	HTTP	194	irc
22	SSH	109	POP2	220	imap3
23	Telnet	110	POP3	389	ldap
25	SMTP	111	sunrpc	443	https

Port	Application	Port	Application	Port	Application
37	time	119	NNTP	520	rip
42	name	123	NTP	554	rtsp
43	whois	143	IMAP	1645/1812	RADIUS
53	domain	144	news	1813	RADIUS accounting
69	TFTP	161	SNMP	1985	hsrp
70	gopher	162	snmptrap		

- TCP flag value, as shown in Table 2-6

Table 2-6 TCP flag values

Flag	Value
URG	0x0020
ACK	0x0010
PSH	0x0008
RST	0x0004
SYN	0x0002
FIN	0x0001

- Packet format (for regular ACLs and VMaps only)
  - Ethernet format (eth2, SNAP, LLC)
  - Ethernet tagging format
  - IP format (IPv4, IPv6)
- Egress port packets (for all ACLs)

Summary of ACL actions

After the packet flows are classified by using ACLs, they can be processed differently. For each ACL, an action can be assigned. The action determines how the switch treats packets that match the classifiers assigned to the ACL. ACL actions include the following actions:

- Pass or Drop the packet
- Remark the packet with a new DiffServ Code Point (DSCP)
- Remark the 802.1p field
- Set the COS queue

ACL order of precedence

When multiple ACLs are assigned to a port, they are evaluated in numeric sequence, based on the ACL number. Lower-numbered ACLs take precedence over higher-numbered ACLs. For example, ACL 1 (if assigned to the port) is evaluated first and has top priority.

If multiple ACLs match the port traffic, only the action of the one with the lowest ACL number is applied. The others are ignored.

If no assigned ACL matches the port traffic, no ACL action is applied.

2.8.7 VLAN maps

A VLAN map (VMAP) is an ACL that can be assigned to a VLAN or VM group rather than to a switch port, as with regular ACLs. A VMAP is useful in a virtualized environment where traffic filtering and metering policies must follow virtual machines (VMs) as they migrate between hypervisors.

Individual VMAP filters are configured in the same fashion as regular ACLs, except that VLANs cannot be specified as a filtering criteria (which is unnecessary, because the VMAP is assigned to a specific VLAN or associated with a VM group VLAN).

2.8.8 Storm-control filters

This feature is available only for Top of Rack IBM System Networking switches.

IBM System Networking RackSwitch provides filters that can limit the number of the following packet types transmitted by switch ports:

- Broadcast packets
- Multicast packets
- Unknown unicast packets (destination lookup failure)

Broadcast storms

Excessive transmission of broadcast or multicast traffic can result in a broadcast storm. A broadcast storm can overwhelm your network with constant broadcast or multicast traffic, and degrade network performance. Common symptoms of a broadcast storm are slow network response times and network operations timing out.

Unicast packets whose destination MAC address is not in the forwarding database are unknown unicasts. When an unknown unicast is encountered, the switch handles it like a broadcast packet and floods it to all other ports in the VLAN (broadcast domain). A high rate of unknown unicast traffic can have the same negative effects as a broadcast storm.

2.9 Quality of Service

You can use Quality of Service (QoS) features to allocate network resources to mission-critical applications at the expense of applications that are less sensitive to such factors as time delays or network congestion. You can configure your network to prioritize specific types of traffic, ensuring that each type receives the appropriate QoS level.

2.9.1 QoS overview

QoS helps you allocate guaranteed bandwidth to the critical applications, and limit bandwidth for less critical applications. Applications such as video and voice must have a certain amount of bandwidth to work correctly; using QoS, you can provide that bandwidth when necessary. Also, you can put a high priority on applications that are sensitive to timing out or that cannot tolerate delay, by assigning their traffic to a high-priority queue.

By assigning QoS levels to traffic flows on your network, you can ensure that network resources are allocated where they are needed most. You can use QoS features to prioritize network traffic, providing better service for selected applications.

Figure 2-22 shows the basic QoS model used by the switch:

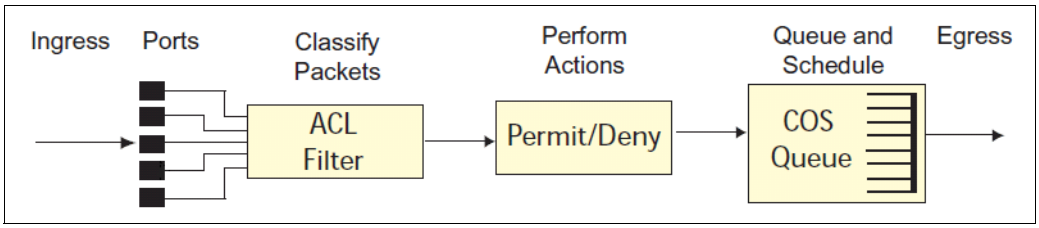

Figure 2-22 QoS model

The basic QoS model works as follows:

- Classify traffic:
  - Read DSCP value.
  - Read 802.1p priority value.
  - Match ACL filter parameters.
- Perform actions:
  - Define bandwidth and burst parameters.
  - Select actions to perform on in-profile and out-of-profile traffic.
  - Deny packets.
  - Permit packets.
  - Mark DSCP or 802.1p priority.
  - Set COS queue (with or without re-marking).
- Queue and schedule traffic:
  - Place packets in one of the COS queues.
  - Schedule transmission based on the COS queue.

2.9.2 Using ACL filters

ACLs are filters that you can use to classify and segment traffic, so you can provide different levels of service to different traffic types. Each filter defines the conditions that must match for inclusion in the filter, and also the actions that are performed when a match is made.

You can use IBM System Networking switches to classify packets based on various parameters. For example:

- Ethernet: Source MAC, destination MAC, VLAN number/mask, Ethernet type, and priority.
- IPv4: Source IP address/mask, destination address/mask, type of service, IP protocol number.
- TCP/UDP: Source port, destination port, and TCP flag.
- Packet format.

For ACL details, see 2.8.6, “Access control lists” on page 97.

2.9.3 Summary of ACL actions

Actions determine how the traffic is treated. The VFSM QoS actions include the following actions:

- Pass or drop the packet.
- Re-mark the packet with a new DiffServ Code Point (DSCP).

- Re-mark the 802.1p field.
- Set the COS queue.

2.9.4 ACL metering and re-marking

You can define a profile for the aggregate traffic that flows through the switch by configuring a QoS meter (if wanted) and assigning ACLs to ports. Actions taken by an ACL are called In-Profile actions. You can configure additional In-Profile and Out-of-Profile actions on a port. Data traffic can be metered, and re-marked to ensure that the traffic flow provides certain levels of service in terms of bandwidth for different types of network traffic.

Metering

QoS metering provides different levels of service to data streams through user-configurable parameters. A meter is used to measure the traffic stream against a traffic profile that you create.

Thus, creating meters yields In-Profile and Out-of-Profile traffic for each ACL, as follows:

- In-Profile: If there is no meter configured or if the packet conforms to the meter, the packet is classified as In-Profile.
- Out-of-Profile: If a meter is configured and the packet does not conform to the meter (exceeds the committed rate or maximum burst rate of the meter), the packet is classified as Out-of-Profile.

Metering: Metering is not supported for IPv6 ACLs. All traffic that matches an IPv6 ACL is considered in-profile for re-marking purposes.

Using meters, you set a Committed Rate in Kbps (in multiples of 64). All traffic within this Committed Rate is In-Profile. Additionally, you can set a Maximum Burst Size that specifies an allowed data burst larger than the Committed Rate for a brief period. These parameters define the In-Profile traffic.

Meters keep the sorted packets within certain parameters. You can configure a meter on an ACL, and perform actions on metered traffic, such as packet re-marking.

Re-Marking

Re-marking allows the treatment of packets to be reset based on new network specifications or wanted levels of service. You can configure the ACL to re-mark a packet as follows:

- Change the DSCP value of a packet, which is used to specify the service level that traffic should receive.
- Change the 802.1p priority of a packet.

2.9.5 DiffServ Code Points

The six most significant bits in the TOS byte of the IP header are defined as DiffServ Code Points (DSCP). Packets are marked with a certain value that depends on the type of treatment the packet must receive in the network device. DSCP is a measure of the QoS level of the packet.

Differentiated Services concepts

To differentiate between traffic flows, packets can be classified by their DSCP value. The Differentiated Services (DS) field in the IP header is an octet, and the first 6 bits, called the DS Code Point (DSCP), can provide QoS functions. Each packet carries its own QoS state in the DSCP. There are 64 possible DSCP values (0-63).

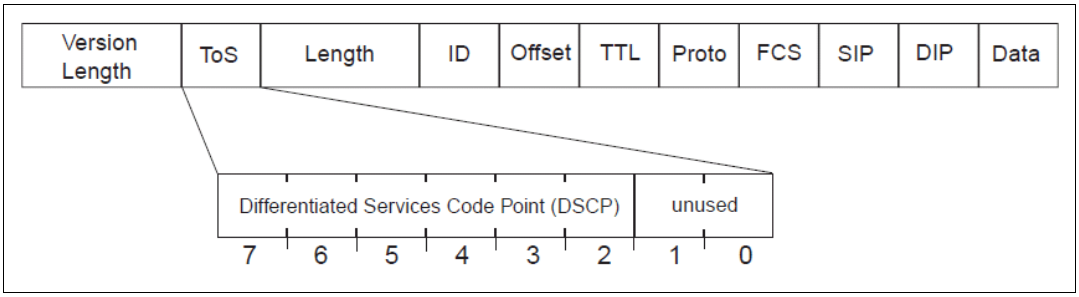

Figure 2-23 IPv4 packet with DSCP field

The switch can perform the following actions on the DSCP:

- Read the DSCP value of ingress packets.
- Re-mark the DSCP value to a new value.
- Map the DSCP value to an 802.1p priority.

After the DSCP value is marked, the switch can use it to direct traffic prioritization.

Per-Hop Behavior

The DSCP value determines the Per-Hop Behavior (PHB) of each packet. The PHB is the forwarding treatment given to packets at each hop. QoS policies are built by applying a set of rules to packets, based on the DSCP value, as they hop through the network.

The default settings are based on the following standard PHBs, as defined in the IEEE standards:

- Expedited Forwarding (EF): This PHB has the highest egress priority and lowest drop precedence level. EF traffic is forwarded ahead of all other traffic. EF PHB is described in RFC 2598, found at:

  http://www.ietf.org/rfc/rfc2598.txt

- Assured Forwarding (AF): This PHB contains four service levels, each with a different drop precedence, as shown in Table 2-7. Routers use drop precedence to determine which packets to discard last when the network becomes congested. AF PHB is described in RFC 2597, found at:

  http://www.ietf.org/rfc/rfc2597.txt

Table 2-7 Assured Forwarding PHB

Drop precedence	Class 1	Class 2	Class 3	Class 4
Low	AF11 (DSCP 10)	AF21 (DSCP 18)	AF31 (DSCP 26)	AF41 (DSCP 34)
Medium	AF12 (DSCP 12)	AF22 (DSCP 20)	AF32 (DSCP 28)	AF42 (DSCP 36)
High	AF13 (DSCP 14)	AF23 (DSCP 22)	AF33 (DSCP 30)	AF43 (DSCP 38)

- Class Selector (CS): This PHB has eight priority classes, with CS7 representing the highest priority, and CS0 representing the lowest priority, as shown in Table 2-8. CS PHB is described in RFC 2474, found at:

  http://www.ietf.org/rfc/rfc2474.txt

Table 2-8 Class Selector PHB

Priority	Class selector	DSCP
Highest	CS7	56
	CS6	48
	CS5	40
	CS4	32
	CS3	24
	CS2	16
	CS1	8
Lowest	CS0	0

QoS levels

Table 2-9 shows the default service levels provided by the VFSM, listed from highest to lowest importance:

Table 2-9 Default QoS service levels

Service level	Default PHB	802.1p priority
Critical	CS7	7
Network Control	CS6	6
Premium	EF, CS5	5
Platinium	AF41, AF42, AF43, and CS4	4
Gold	AF31, AF32, AF33, and CS3	3
Silver	AF21, AF22, AF23, and C22	2
Bronze	AF11, AF12, AF13, and CS1	1
Standard	DF and CS0	0

DSCP re-marking and mapping

The switch can use the DSCP value of ingress packets to re-mark the DSCP to a new value, and to set an 802.1p priority value.

2.9.6 QoS 802.1p

IBM Networking OS provides Quality of Service functions based on the priority bits in a packet's VLAN header. (The priority bits are defined by the 802.1p standard within the IEEE 802.1q VLAN header.) The 802.1p bits, if present in the packet, specify the priority that should be given to packets during forwarding. Packets with a numerically higher (non-zero) priority are given forwarding preference over packets with lower priority bit value.

The IEEE 802.1p standard uses eight levels of priority (0 - 7). Priority 7 is assigned to highest priority network traffic, such as OSPF or RIP routing table updates, priorities 5 - 6 are assigned to delay-sensitive applications, such as voice and video, and lower priorities are assigned to standard applications. A value of 0 (zero) indicates a "best effort" traffic prioritization, and this value is the default when traffic priority is not configured on your network. The VFSM can filter packets based on the 802.1p values, and it can assign or overwrite the 802.1p value in the packet.

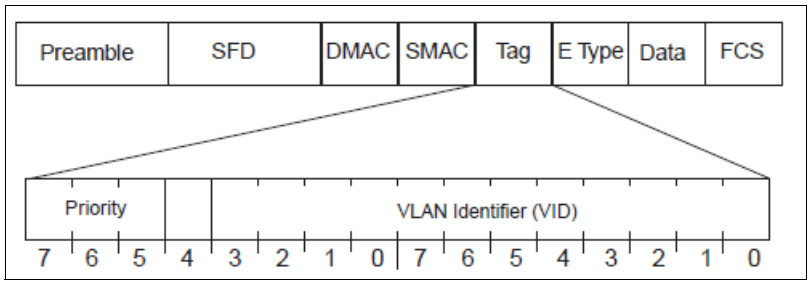

Figure 2-24 Layer 2 802.1q/802.1p tagged frame

Ingress packets receive a priority value, as follows:

- Tagged packets: The switch reads the 802.1p priority in the VLAN tag.
- Untagged packets: The switch tags the packet and assigns an 802.1p priority value, based on the port's default 802.1p priority.

Egress packets are placed in a COS queue based on the priority value, and scheduled for transmission based on the COS queue number. Higher COS queue numbers provide forwarding precedence.

2.9.7 Queuing and scheduling

IBM System Networking switches can be configured to have either two or eight output Class of Service (COS) queues per port, into which each packet is placed. Each packet's 802.1p priority determines its COS queue, except when an ACL action sets the COS queue of the packet.

You can configure the following attributes for COS queues:

- Map 802.1p priority value to a COS queue.
- Define the scheduling weight of each COS queue.

The scheduling weight can be set 0 - 15. Weight values 1 - 15 set the queue to use weighted round-robin (WRR) scheduling, which distributes larger numbers of packets to queues with the highest weight values. For distribution purposes, each packet is counted the same, regardless of the packet's size.

A scheduling weight of 0 (zero) indicates strict priority. Traffic in the strict priority queue has precedence over other all queues. If more than one queue is assigned a weight of 0, the strict queue with highest queue number is served first. After all traffic in strict queues is delivered, any remaining bandwidth is allocated to the WRR queues, divided according to their weight values.

Strict scheduling: Use caution when assigning strict scheduling to queues. Heavy traffic in queues assigned with a weight of 0 can starve lower priority queues.

3

Reference architectures

This chapter presents the network architecture used in this book to implement a 10 Gb Ethernet solution by using IBM System Networking switches. The design is based on preferred practices and the experience of the authors and is meant to provide support for the implementation chapters later in this book.

The network topology and configuration details presented in the following sections are referenced and used as input for equipment configuration and verification examples to show the software features implementation steps and guidelines.

The network solution implemented in this book is not a complete data center architecture suited for a production environment, but a simple topology used to show configuration examples for features, commonly used in a real world implementation.

A reader that has the same (or equivalent) equipment used in the reference architectures at their disposal should be able to replicate the configurations by following the steps described in the implementation chapters and arrive at the same result.

The book focuses on the network side configuration, but for testing purposes, sample hosts are used to show end-to-end connectivity as a result of the implementation.

© Copyright IBM Corp. 2012. All rights reserved.

3.1 Overview of the reference architectures

The reference architectures describe a mixed environment of both stand-alone and IBM BladeCenter embedded switches that are integrated in a fully functional network. These switches are able to provide end-to-end communication in a data center, for servers that run different operating systems (Windows and Linux) and IP protocol versions (IPv4 and IPv6).

We follow the three-tier data center design that focuses on the access and aggregation layers, because those layers are the layers IBM System Networking products use.

Both designs differ in how the Access Layer is implemented:

- The Top-of-Rack architecture uses a pair of IBM System Networking stand-alone switches as the access layer.
- The BladeCenter architecture uses a pair of IBM Virtual Fabric 10Gb Switch Modules in an IBM BladeCenter chassis as the access layer.

The architectures share a common aggregation layer.

The following equipment is used for the network topology described in this chapter:

- Two RackSwitches G8264 for the shared aggregation layer
- Two RackSwitches G8124 for the access layer in the Top-of-Rack architecture
- Two IBM Virtual Fabric 10Gb Switch Modules for the access layer in the BladeCenter architecture
- One IBM System x3550 M3 running Windows Server 2008R2 for the host connected to the RackSwitch access layer
- One IBM BladeCenter HS22 running Red Hat Enterprise Linux for the BladeCenter host

3.2 Top-of-Rack architecture

A Top-of-Rack architecture describes the aggregation and access layers topology by using stand-alone switches. It has the following components:

- Two RackSwitches G8264 were used for the aggregation layer. They are redundantly connected with 40 Gbps links to each other and 10 Gbps to the access layer switches.
- Two RackSwitches G8124 were used for the access layer. They are redundantly connected with 10 Gbps links to each other and aggregation layer switches.
- The server has two 10 Gb uplinks connecting to each access layer switch.

Figure 3-1 shows an overview of the Top-of-Rack architecture.

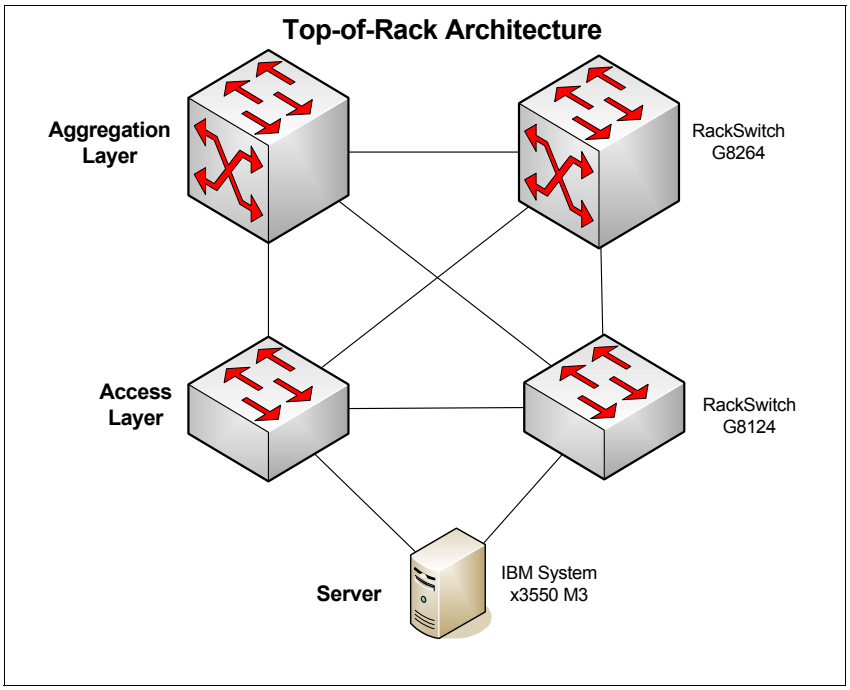

Figure 3-1 Top-of-Rack architecture overview

3.2.1 Layer 1 architecture

The Layer 1, or physical layer, architecture includes the hardware components, physical cabling, connectors, interfaces, and host names.

Hardware components

The list of equipment used in this topology is shown in Table 3-1.

Table 3-1 Top-of-Rack - hardware

Host name	Model	Description
SRV-1	System x3550 M3	
ACC-1	RackSwitch G8124	Access Layer Switch 1
ACC-2	RackSwitch G8124	Access Layer Switch 2
AGG-1	RackSwitch G8264	Aggregation Layer Switch 1
AGG-2	RackSwitch G8164	Aggregation Layer Switch 2

Table 3-2 shows the switch interfaces used for connecting the devices. The port numbering assumes that aggregation (RackSwitch G8264) is in QSFP+ 40GbE mode.

Table 3-2 Top-of-Rack - interconnections

Device #1	Interface device #1	Device #2	Interface Device #2	Interface type	Connector/ cable type
SRV-1	NIC0	ACC-1	port7	10 Gbps	10G SFP+ Transceiver, fiber
SRV-1	NIC1	ACC-2	port7	10 Gbps	10G SFP+ Transceiver, fiber
ACC-1	port1	AGG-1	port17	10 Gbps	DAC, copper
ACC-1	port2	AGG-1	port18	10 Gbps	DAC, copper
ACC-1	port3	AGG-2	port19	10 Gbps	DAC, copper
ACC-1	port4	AGG-2	port20	10 Gbps	DAC, copper
ACC-1	port5	ACC-2	port5	10 Gbps	DAC, copper
ACC-1	port6	ACC-2	port6	10 Gbps	DAC, copper
ACC-2	port1	AGG-2	port17	10 Gbps	DAC, copper
ACC-2	port2	AGG-2	port18	10 Gbps	DAC, copper
ACC-2	port3	AGG-1	port19	10 Gbps	DAC, copper
ACC-2	port4	AGG-1	port20	10 Gbps	10G SFP+ Transceiver, fiber
ACC-2	port5	ACC-1	port5	10 Gbps	DAC, copper
ACC-2	port6	ACC-1	port6	10 Gbps	DAC, copper
AGG-1	port1	AGG-2	port1	40 Gbps	DAC, copper (QSFP+)
AGG-1	port5	AGG-2	port5	40 Gbps	DAC, copper (QSFP+)
AGG-2	port1	AGG-1	port1	40 Gbps	DAC, copper (QSFP+)
AGG-2	port5	AGG-1	port5	40 Gbps	DAC, copper (QSFP+)

Figure 3-2 shows a Layer 1 architecture with the host names and interfaces used for inter-switch connections.

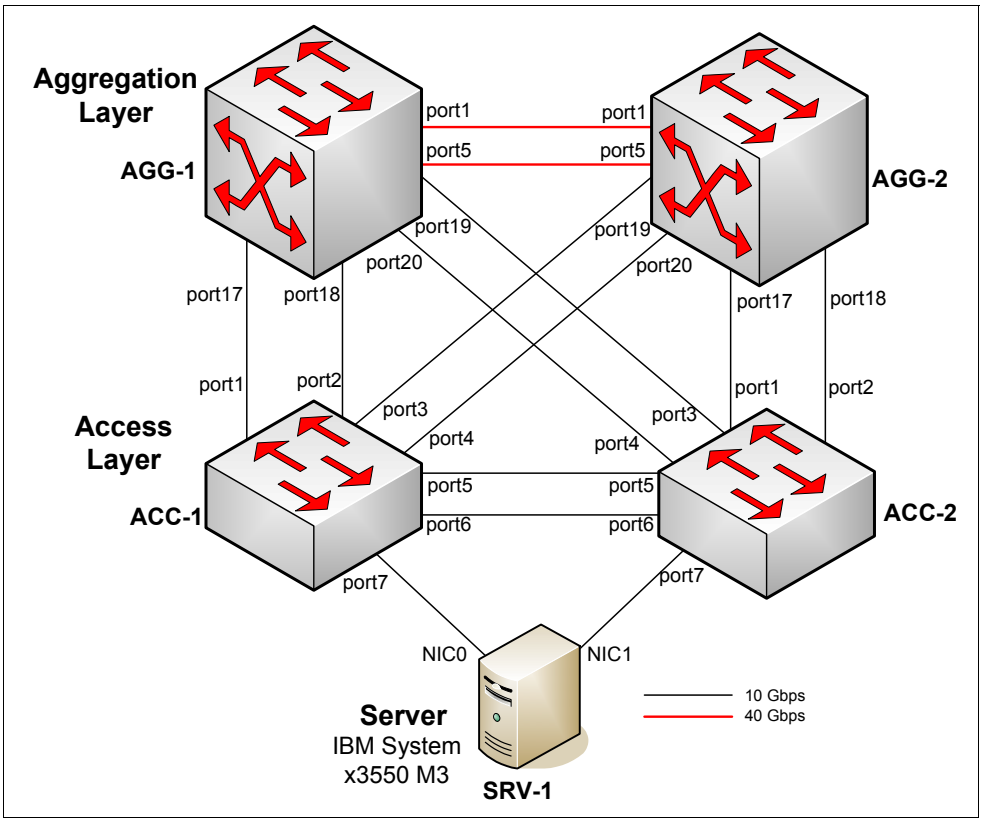

Figure 3-2 Top-of-Rack - Layer 1 architecture

3.2.2 Layer 2 architecture

Top-of-Rack architecture brings routing down to the access layer, which means that Layer 2, or the Spanning Tree Protocol (STP) domain, terminates at access layer switches (ACC-1 and ACC-2).

Virtual Local Area Networks

Table 3-3 shows Virtual Local Area Networks (VLANs) used in the topology and the member ports of those VLANs.

Table 3-3 VLANs and port assignment

VLAN number	VLAN name	Description	Member ports
VLAN4095	Mgmt VLAN	Management VLAN	ACC-1, port127 (untagged) ACC-2, port127(untagged)
VLAN10	Server 1 VLAN	VLAN for Server farm 1	ACC-1, port5 (tagged) ACC-1, port6 (tagged) ACC-1, port7 (untagged) ACC-2, port5 (tagged) ACC-2, port6 (tagged) ACC-2, port7 (untagged)

VLAN number	VLAN name	Description	Member ports
VLAN101	ACC-1 - AGG-1	ACC-1 AGG1 p2p network	ACC-1, port1 (untagged) ACC-1, port2 (untagged) AGG-1, port17 (untagged) AGG-1, port18 (untagged)
VLAN102	ACC-1 - AGG-2	ACC-1 - AGG-2 p2p network	ACC-1, port3 (untagged) ACC-1, port4 (untagged) AGG-2, port19 (untagged) AGG-2, port20 (untagged)
VLAN103	ACC-2 - AGG-1	ACC-2 - AGG-1 p2p network	ACC-2, port3 (untagged) ACC-2, port4 (untagged) AGG-1, port19 (untagged) AGG-1, port20 (untagged)
VLAN104	ACC-2 - AGG-2	ACC-2 - AGG-2 p2p network	ACC-2, port1 (untagged) ACC-2, port2 (untagged) AGG-2, port17 (untagged) AGG-2, port18 (untagged)
VLAN100	AGG-1 - AGG-2	AGG-1 AGG-2 p2p network	AGG-1, port1 (tagged) AGG-1, port5 (tagged) AGG-2, port1 (tagged) AGG-2, port5 (tagged)

Spanning Tree Protocol

The STP mode that we use in our topology is Per-VLAN Rapid Spanning Tree Protocol (PVRSTP) with the bridge priorities shown in Table 3-4.

Table 3-4 Spanning Tree Protocol bridge priorities

VLAN	Priority on ACC-1	Priority on ACC-2	Priority on AGG-1	Priority on AGG-1
10	0	4096	There is no VLAN10 on the switch.	There is no VLAN10 on the switch.

For other VLANs, the priorities are left at the default values, leaving the root bridge selection process to rely on the MAC addresses of the switches.

Ports 7 on access switches (ACC-1 and ACC-2) are configured as Fast Forward ports.

To prevent unwanted topology changes, caused by plugging another switch into the topology, configure BPDUGuard on all ports on ACC-1 and ACC-2 switches other than the ports used in the topology.

Trunks

We group the inter-switch connection into trunks, as shown in Table 3-5.

Table 3-5 Trunk configuration

Switch	Trunk	Trunk members	Static or LACP
ACC-1	portchannel3	port5, port6	LACP
ACC-2	portchannel3	port5, port6	LACP
ACC-1	portchannel1	port1, port2	static
ACC-1	portchannel2	port3, port4	static
ACC-2	portchannel1	port1, port2	static
ACC-2	portchannel2	port3, port4	static
AGG-1	portchannel1	port17, port18	static
AGG-1	portchannel2	port19, port20	static
AGG-1	portchannel3	port1, port5	static
AGG-2	portchannel1	port17, port18	static
AGG-2	portchannel2	port19, port20	static
AGG-2	portchannel3	port1, port5	static

3.2.3 Layer 3 architecture

The Layer 3 architecture describes the IP addressing used and the initial set-up for Layer 3 protocols such as Virtual Router Redundancy Protocol (VRRP) and Open Shortest Path First (OSPF).

IP addressing

The IPv4 and IPv6 addressing that we used in the topology is presented in Table 3-6.

Table 3-6 IP addressing Top-of-Rack network

VLAN	IPv4	IPv6	Description
VLAN 4095	172.25.0.0/16	n/a	Management network
VLAN 10	10.0.10.0/24	FC10::0/64	VLAN10
VLAN101	10.0.101.0/30	FC11::0/64	Point-to-point link subnet between ACC-1 and AGG-1
VLAN102	10.0.102.0/30	FC12::0/64	Point-to-point link subnet between ACC-1 and AGG-2
VLAN103	10.0.103.0/30	FC13::0/64	Point-to-point link subnet between ACC-2 and AGG-1

VLAN	IPv4	IPv6	Description
VLAN104	10.0.104.0/30	FC14::0/64	Point-to-point link subnet between ACC-2 and AGG-2
VLAN100	10.0.100.0/30	FC00::0/64	Point-to-point link subnet between AGG-1 and AGG-2

In Table 3-7, we show the management IPv4 addresses, that is, addresses assigned to the management interfaces.

Table 3-7 Management IP addresses

Switch	Management IP address	Management interface
ACC-1	172.25.101.122	port25 (MGTA)
ACC-2	172.25.101.123	port25 (MGTA)
AGG-1	172.25.101.120	port65 (MGT)
AGG-2	172.25.101.121	port65 (MGT)

Table 3-8 shows the IPv4 and IPv6 host addresses assignment for the devices.

Table 3-8 IPv4 and IPv6 host addresses assignment

VLAN	IPv4 interface	IPv4 address	IPv6 interface	IPv6 address	Device
VLAN10	Team Adapter	10.0.10.10	Team Adapter	FC10::10/64	SRV-1
VLAN10	10	10.0.10.2	106	FC10::2/64 FC10::1/64 secondary anycast	ACC-1
VLAN10	10	10.0.10.3	106	FC10::3/64 FC10::1/64 secondary anycast	ACC-2
VLAN10	Virtual router 10	10.0.10.1	n/a	n/a	VRRP 10
VLAN101	101	10.0.101.1	111	FC11::1/64	ACC-1
VLAN102	102	10.0.102.1	112	FC12::1/64	ACC-1
VLAN104	104	10.0.104.2	114	FC14::2/64	ACC-2
VLAN103	103	10.0.103.2	113	FC13::2/64	ACC-2
VLAN101	101	10.0.101.2	111	FC11::2/64	AGG-1
VLAN103	103	10.0.103.1	113	FC13::1/64	AGG-1
VLAN100	100	10.0.100.1	110	FC00::1/64	AGG-1
VLAN104	104	10.0.104.1	114	FC14::1/64	AGG-2

VLAN	IPv4 interface	IPv4 address	IPv6 interface	IPv6 address	Device
VLAN102	102	10.0.102.2	112	FC12::2/64	AGG-2
VLAN100	100	10.0.100.2	110	FC00::2/64	AGG-2

Loopback interfaces are used for static router-ID assignment to use with OSPFv2 and OSPFv3. Table 3-9 shows the loopback interfaces IP addresses assignment

Table 3-9 Loopback interfaces

IP address	Loopback	Device
1.1.1.1/32	Loopback 1	AGG-1
1.1.1.2/32	Loopback 1	AGG-2
2.2.2.1/32	Loopback 1	ACC-1
2.2.2.2/32	Loopback 1	ACC-2

Keep the STP bridge priorities consistent with VRRP priorities, that is, make sure that the switch that is a root bridge for STP is also an active router for VRRP for that VLAN.

Table 3-10 shows the VRRP groups and priorities what we used in the topology.

Table 3-10 VRRP groups and priorities

VRRP group	ACC-1 priority	ACC-2 priority
10	105	100 (default)

OSPFv2 and OSPFv3

We run single-area Open Shortest Path First (OSPF) in area number 0 (backbone).

OSPF is run on all Layer 3 interfaces. IP interfaces of VLANs are configured with the *passive-interface* option, so that the hellos are not exchanged over them.

Figure 3-3 shows the Layer 3 architecture of our Top-of-Rack design.

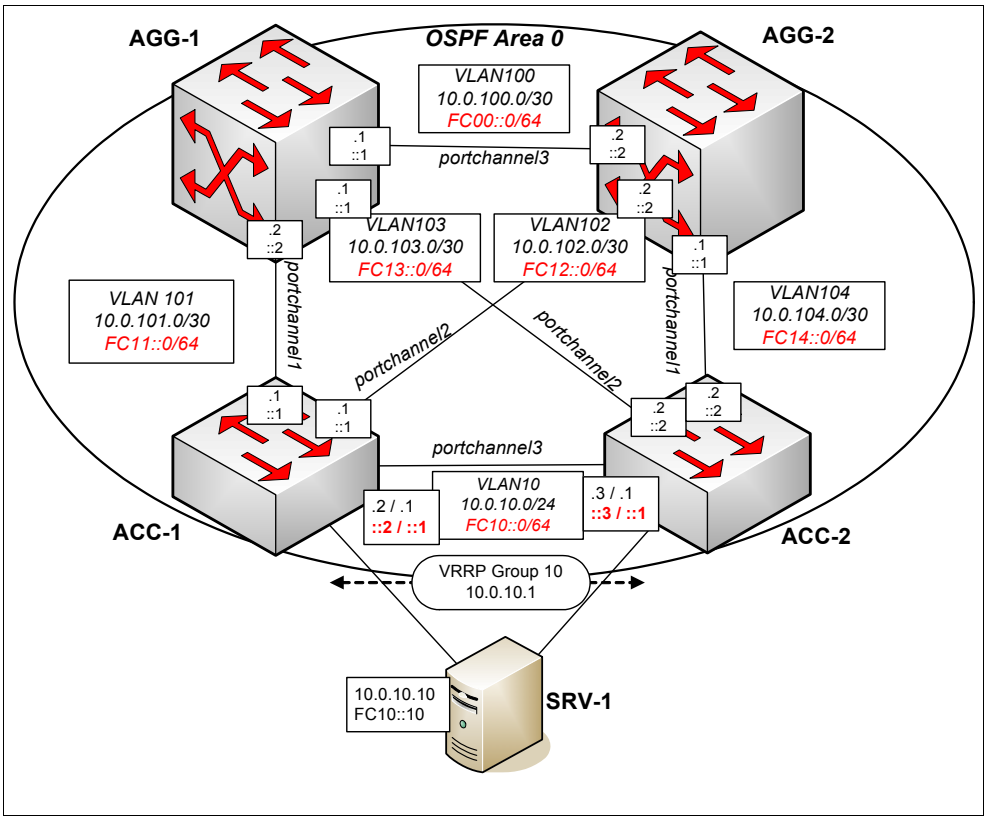

Figure 3-3 Top-of-Rack - Layer 3 architecture

3.3 IBM BladeCenter architecture

In our BladeCenter architecture, an access layer is made up of two IBM Virtual Fabric 10Gb Switch Modules in a BladeCenter chassis. The switches are configured as a stack and the servers are blade servers. The aggregation layer of BladeCenter design is used with Top-of-Rack design.

Figure 3-4 shows the components of our BladeCenter architecture.

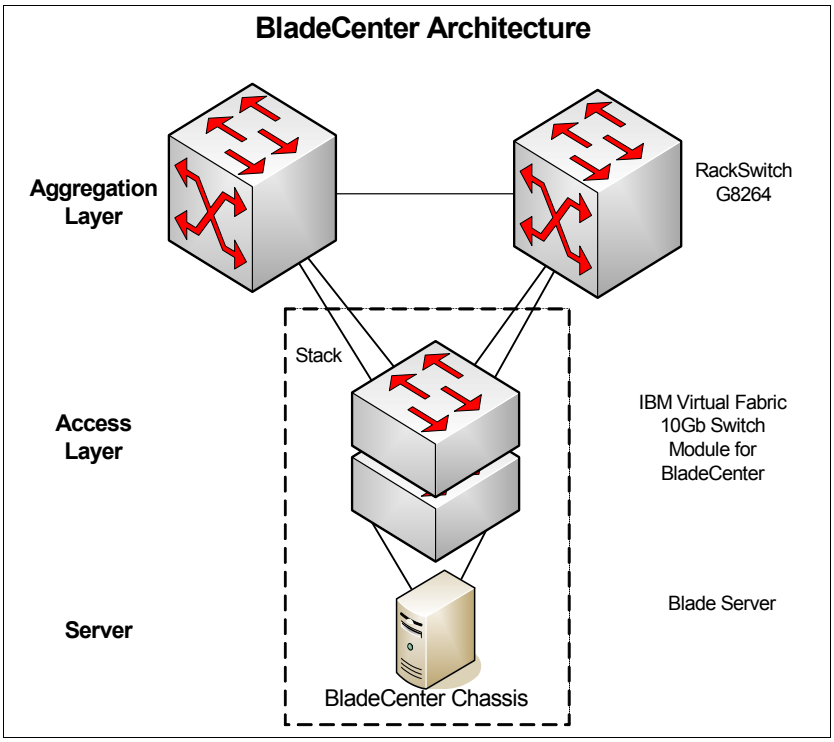

Figure 3-4 BladeCenter - architecture overview

3.3.1 Layer 1 architecture

Layer 1, or physical layer, architecture includes the hardware components, physical cabling, connectors, interfaces, and host names.

Hardware components

Table 3-11 shows devices used in our topology.

Table 3-11 BladeCenter architecture hardware

Host name	Model	Description
SRV-3	IBM BladeCenter HS22	Blade Server
ACC-3-1	IBM BladeCenter Virtual Fabric 10Gb Switch Module	Access Layer Switch 3
ACC-3-2	IBM BladeCenter Virtual Fabric 10Gb Switch Module	Access Layer Switch 4
AGG-1	RackSwitch G8264	Aggregation Layer Switch 1
AGG-2	RackSwitch G8264	Aggregation Layer Switch 2

Figure 3-5 shows the chassis of the IBM BladeCenter H used in the access layer of our BladeCenter architecture.

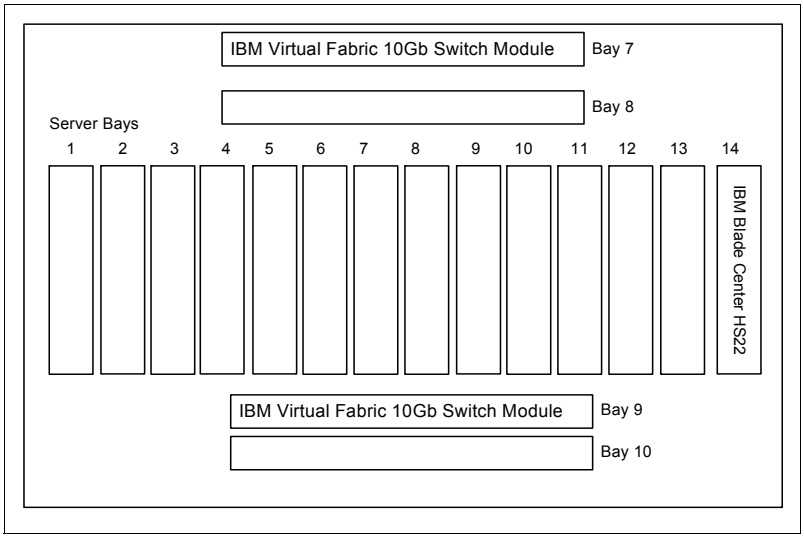
IBM Virtual Fabric 10Gb Switch Module
Bay 7
Bay 8
Server Bays
1 2 3 4 5 6 7 8 9 10 11 12 13 14
IBM Blade Center HS22
IBM Virtual Fabric 10Gb Switch Module
Bay 9
Bay 10

Figure 3-5 IBM BladeCenter H chassis used in the access layer of our BladeCenter architecture

The physical connections details used for the reference architecture installation are summarized in Table 3-12.

Table 3-12 BladeCenter - interconnections

Device #1	Interface device #1	Device #2	Interface device #2	Interface type	Connector/Cable type
SRV-3	NIC2	ACC-3	port1:14	10 Gbps	Internal bus
SRV-3	NIC3	ACC-3-2	port2:14	10 Gbps	Internal bus
ACC-3	port1:17	AGG-1	port21	10 Gbps	10G SFP+ Transceiver, fiber
ACC-3	port1:18	AGG-2	port21	10 Gbps	10G SFP+ Transceiver, fiber
ACC-3	port1:25	ACC-3-2	port2:25	10 Gbps	DAC copper
ACC-3	port1:26	ACC-3-2	port2:26	10 Gbps	DAC copper
ACC-3-2	port2:17	AGG-1	port22	10 Gbps	10G SFP+ Transceiver, fiber
ACC-3-2	port2:18	AGG-2	port22	10 Gbps	10G SFP+ Transceiver, fiber
ACC-3-2	port2:25	ACC-3	port1:25	10 Gbps	DAC copper
ACC-3-2	port2:26	ACC-3	port1:26	10 Gbps	DAC copper
AGG-1	port1	AGG-2	port1	40 Gbps	DAC copper (QSFP+)
AGG-1	port5	AGG-2	port5	40 Gbps	DAC copper (QSFP+)

Device #1	Interface device #1	Device #2	Interface device #2	Interface type	Connector/Cable type
AGG-2	port1	AGG-1	port1	40 Gbps	DAC copper (QSFP+)
AGG-2	port5	AGG-1	port5	40 Gbps	DAC copper (QSFP+)

Figure 3-6 shows the Layer 1 architecture based on the information in Table 3-12 on page 118.

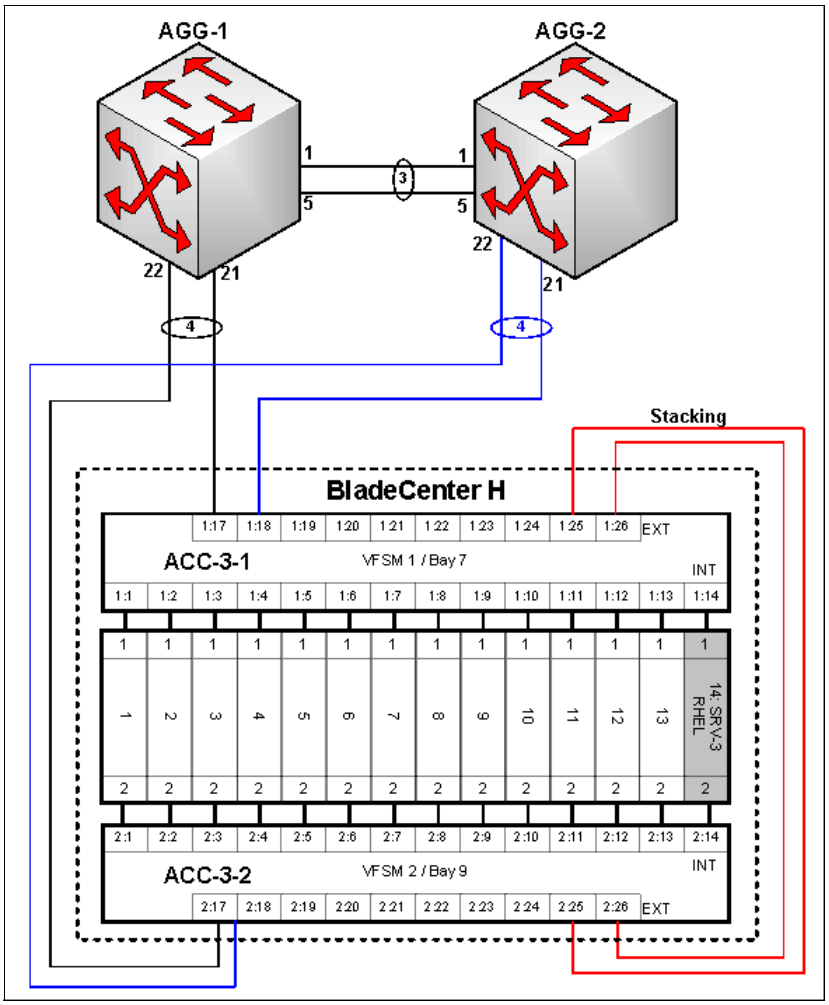

Figure 3-6 BladeCenter - Layer 1 architecture

3.3.2 Layer 2 architecture

Access layer switches (ACC-3 and ACC-4) are stacked using a pair of connections on external ports 9 and 10 (port25 and port26). ACC-3 is configured as a stack master.

The BladeCenter architecture is based on Per-VLAN Rapid Spanning Tree Protocol (PVRSTP) running at all layers up to the aggregation layer.

VLANs

Table 3-13 shows VLANs used in the topology and the member ports of those VLANs.

Table 3-13 VLANs and port assignment

VLAN number	VLAN name	Description	Member ports
VLAN4095	Mgmt VLAN	Management VLAN	ACC-3, port15 (MGT1) ACC-3-2, port15 (MGT1)
VLAN30	Server 3 VLAN	VLAN for Server farm 3	ACC-3, port 1:14 (untagged) ACC-3, port 1:17 (tagged) ACC-3, port 1:18 (tagged) ACC-3-2, port 2:14 (untagged) ACC-3-2, port 2:17 (tagged) ACC-3-2, port 2:18 (tagged) AGG-1, port1 (tagged) AGG-1, port5 (tagged) AGG-1, port21 (tagged) AGG-1, port22 (tagged) AGG-2, port1 (tagged) AGG-2, port5 (tagged) AGG-2, port21 (tagged) AGG-2, port22 (tagged)

Spanning Tree Protocol

The STP mode that we use in our topology is Per-VLAN Rapid Spanning Tree Protocol (PVRSTP). Table 3-14 shows the STP priorities for our different VLANs.

Table 3-14 STP bridge priorities

VLAN (STG)	Priority on ACC-3	Priority on AGG-1	Priority on AGG-2
30 (30)	Default (61470)	0	4096

Port1:14 and port2:14 on access switches (ACC-3-1 and ACC-3-2) are configured as Fast Forward ports.

To prevent unwanted topology changes, caused by plugging another switch into topology, configure BPDU Guard on all ports on ACC-3-1 and ACC-3-2 switches other than those ports used in the topology.

3.3.3 Layer 3 architecture

This section describes the IPv4 and IPv6 addressing plan and the complete Layer 3 architecture that are part of both Top-of-Rack and BladeCenter implementations.

IPv4 and IPv6 addressing

Table 3-15 shows the IPv4 and IPv6 address spaces assigned for VLAN 30.

Table 3-15 IP address ranges

VLAN	IPv4	IPv6	Description
30	10.0.30.0/24	FC30::0/64	VLAN30

Table 3-16 shows the IP addresses assigned to devices that have IP interfaces in VLAN30 and VLAN40.

Table 3-16 IPv4 and IPv6 addresses assignment

VLAN	IPv4 interface	IPv4 address	IPv6 interface	IPv6 address	Device
VLAN30	Bonding interface	10.0.30.30	Bonding interface	FC30::30/64	SRV-3
VLAN30	30	10.0.30.2	36	FC30::2/64 FC30::1/64 secondary anycast	AGG-1
VLAN30	30	10.0.30.3	36	FC30::3/64 FC30::1/64 secondary anycast	AGG-2
VLAN30	Virtual router 30	10.0.30.1	n/a	n/a	VRRP 30 on AGG-1 / AGG-2

VRRP configuration

The communications within one VLAN is done by using the access layer switches. Inter-VLAN communications is done by using aggregation layer switches (AGG-1, AGG2) configured with VRRP groups for different VLANs.

The VRRP priorities should be kept in-line with STP bridge priorities, that is, the STP root bridge should be the VRRP active router. Table 3-17 shows the VRRP priorities.

Table 3-17 VRRP groups and priorities

VRRP group	AGG-1 priority	AGG-2 priority
30	105	100 (default)

3.4 Final architecture

Figure 3-7 shows the final network topology, with the connected hosts that are able to communicate with each other on IPv4 and IPv6.

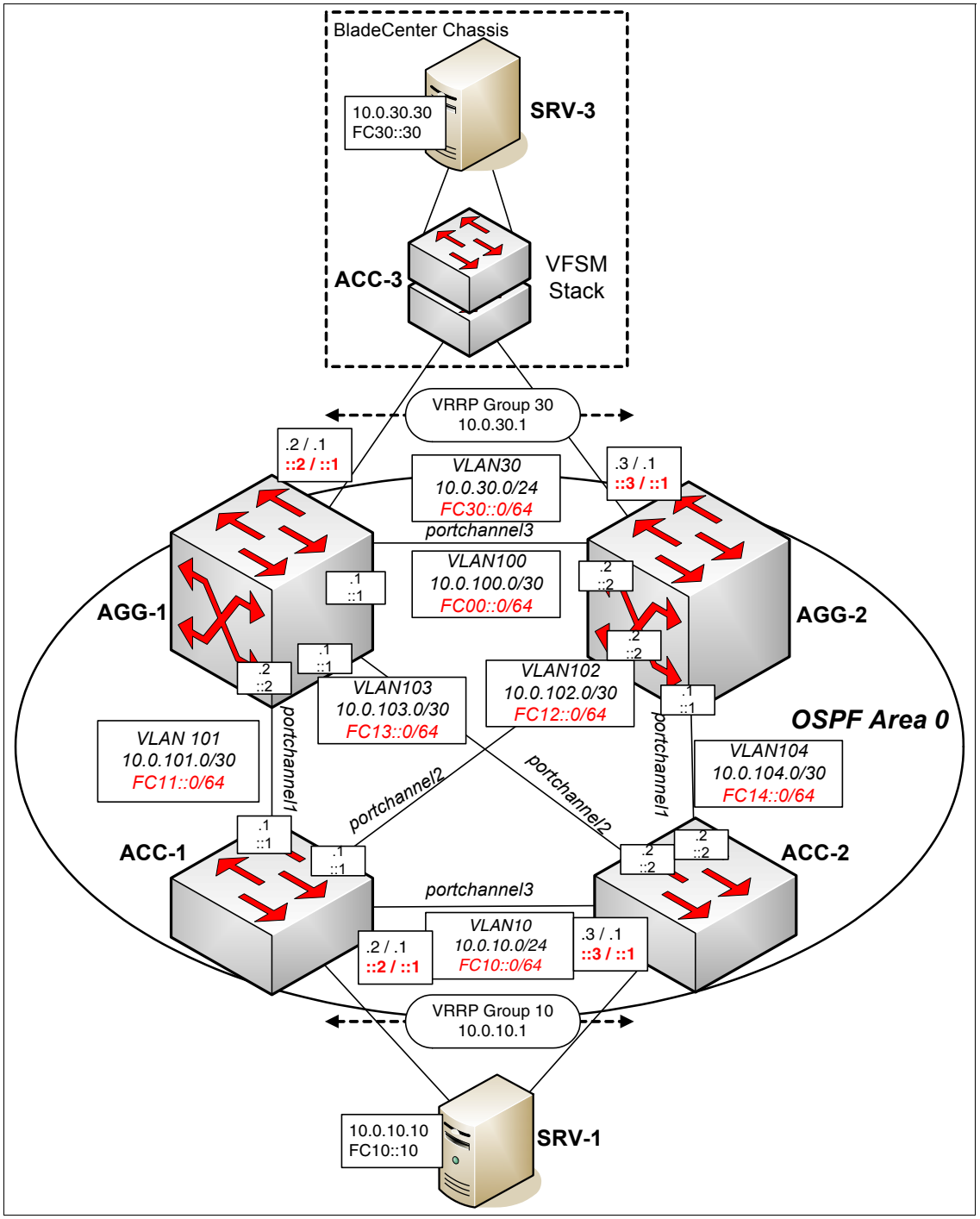

Figure 3-7 Final architecture

4

Initial configuration: IBM System Networking 10Gb Ethernet switches

In this chapter, we describe the steps to be performed for the initial configuration of the IBM System Networking 10Gb Ethernet switches, both in the Top-of-Rack (TOR) RackSwitch G8264 and embedded switch IBM 10Gb Virtual Fabric Module for BladeCenter versions. We also provide an introduction to the IBM System Networking Element Manager (Previously know as BLADE Harmony Manager) software.

For detailed hardware planning information, and the hardware installation procedures, see the appropriate installation guide listed in “Related publications” on page 333.

Terminology: Because of the recent acquisition of BNT by IBM, some of the product names used in this publication may change.

© Copyright IBM Corp. 2012. All rights reserved.

4.1 Overview of the initial setup

The steps cover the following elements for the hardware:

- Terminal connection
- Setting up the IP address of the switch
- Configuring the date and time
- Security

For the IBM System Networking Element Manager software, we describe how to:

- Install it
- Configure the basic options

Switch initial setup

Every time we receive a new switch and we want to install it in our network, there is a set of basic initial configuration tasks that should be done. These include setting the date and time, changing the administrator's password, and some other basic ones. And, to perform these tasks, we should first connect to the switch. In the following sections, we describe how to perform this connection and these tasks.

4.2 Administration interfaces

The RackSwitch G8264 and the IBM Virtual Fabric 10Gb Switch Module have many management interfaces where you can perform the initial setup tasks. In this section, we cover the common methods for both switches, and also the ones that are specific to the BladeCenter module.

Both switches are able to perform advanced tasks right away. However, some basic setup is needed when you first use the switch.

IBM Networking OS, the operating system that runs inside the switches, provides three main interfaces for administration purposes. These interfaces are:

- A text-based command-line interface (CLI) and menu system for remote access with Telnet and SSH sessions
- The Browser-Based interface (BBI), which can be used with a standard web browser
- SNMP support for access with network management software, such as IBM Systems Director and many others

For the IBM Virtual Fabric 10Gb Switch Module for BladeCenter, there is one additional interface that can be used: The BladeCenter Advance Management Module (AMM) interface, which is used also for general management of the chassis.

Management interfaces: To access the management interfaces, you must know the IP address of the management interface of the switch. If the switch is brand new, and you are working on it for the first time, you must configure this IP address. In the rest of this chapter, we describe this process in more detail.

The default user for administration purposes is admin, and the default password is admin as well. We describe how to change these settings in "User management" on page 137.

4.2.1 Console, Telnet, and Secure Shell (SSH)

The IBM Networking OS CLI provides a simple and direct method for switch administration. Using a basic terminal, you have an organized hierarchy of menus, each with logically related submenus and commands. You can use these items to view detailed information and statistics about the switch, and to perform any necessary configuration and switch software maintenance.

This text-mode console access can be done through a direct attached serial cable to the switch, an SSH terminal session, or Telnet.

Switch administration: The factory default settings permit initial switch administration through only the built-in serial port. All other forms of access require additional switch configuration before they can be used.

4.2.2 Browser-Based interface

The web interface, also called the Browser-Based interface (BBI), for the switches is available on the IP address of the management port of the switch, on port 80. To open the BBI, open your web browser and point to the IP address of the management port. By default, port 80 is the one used for HTTP, so you do not need to specify it in the web browser navigation bar.

In our example, one of our switches is on IP address `172.25.101.7`. This switch is the IBM Virtual Fabric 10Gb Switch Module embedded switch in a BladeCenter H chassis. The results should be similar if you use a TOR model.

To access the interface, access the IP address of your switch from the web page. We use the IP address `http://172.25.101.7` (Figure 4-1). You are prompted to log on. In this case, we use the default login with user admin and password admin.

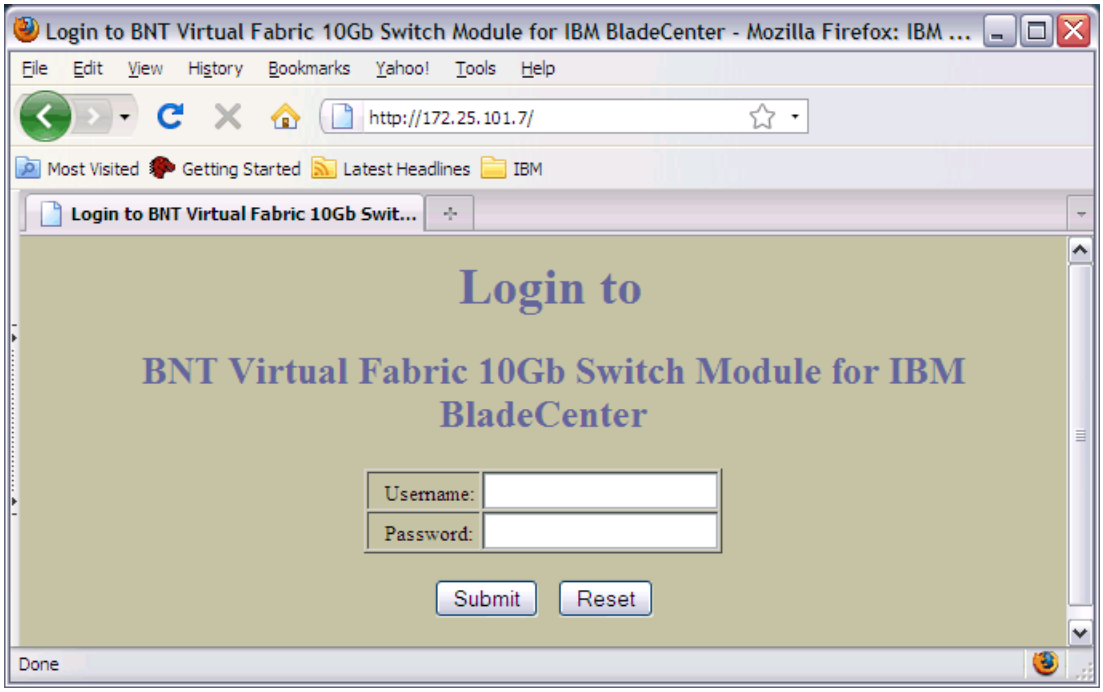
Login to BNT Virtual Fabric 10Gb Switch Module for IBM BladeCenter - Mozilla Firefox: IBM ...
File Edit View History Bookmarks Yahoo! Tools Help
http://172.25.101.7/
Most Visited Getting Started Latest Headlines IBM
Login to BNT Virtual Fabric 10Gb Swit...
Login to
BNT Virtual Fabric 10Gb Switch Module for IBM BladeCenter
Username:
Password:
Submit
Reset
Done

Figure 4-1 Web interface for IBM Networking OS

After successfully logging on, the Switch Dashboard is displayed, as shown in Figure 4-2.

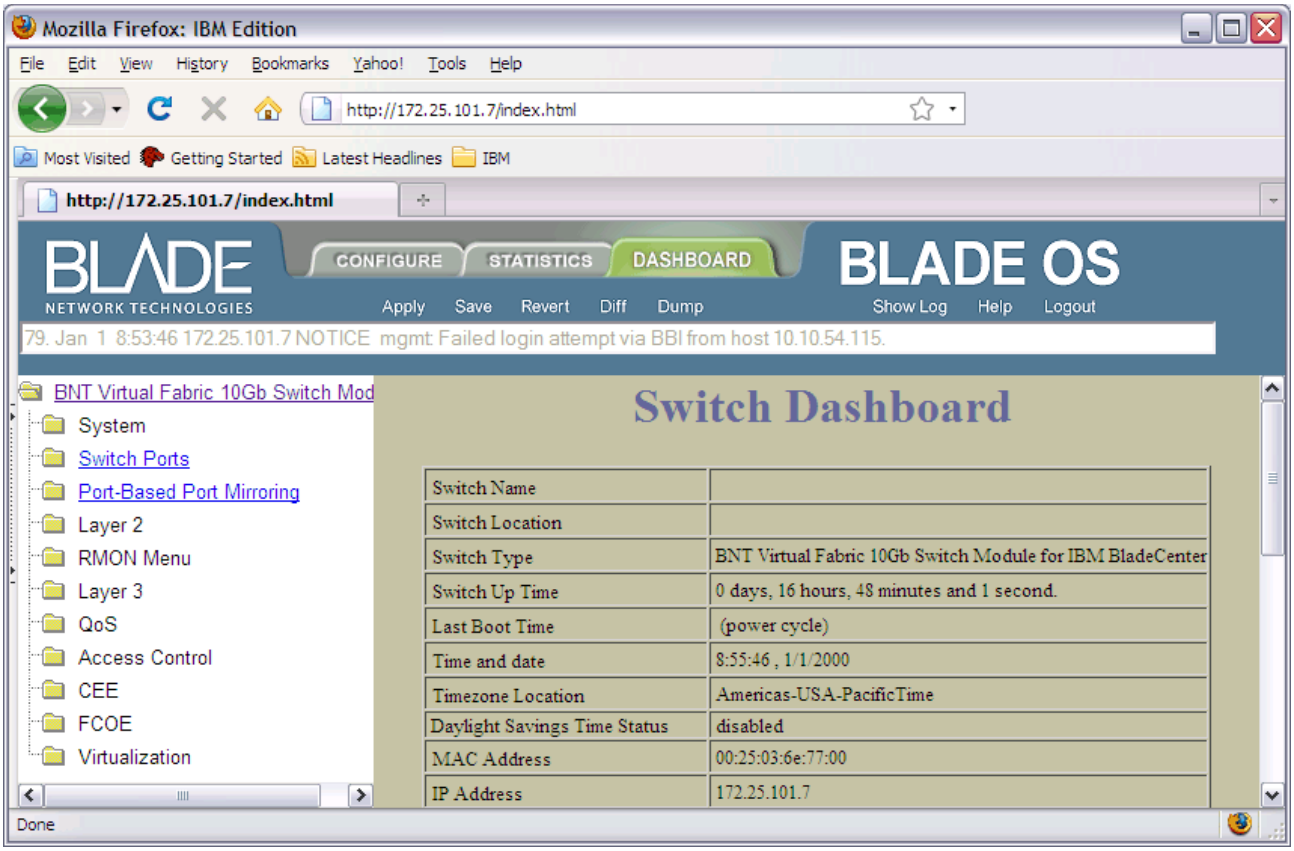

Figure 4-2 Switch Dashboard

For more details about the web interface, see following documents:

- IBM System Networking RackSwitch G8264 Browser-Based Interface Quick Guide:
  https://www-304.ibm.com/support/docview.wss?uid=isg3T7000296&aid=1
- IBM System Networking RackSwitch G8124 Browser-Based Interface Quick Guide:
  https://www-304.ibm.com/support/docview.wss?uid=isg3T7000301&aid=1
- IBM System Networking RackSwitch G8052 Browser-Based Interface Quick Guide:
  https://www-304.ibm.com/support/docview.wss?uid=isg3T7000348&aid=1

During the first boot of your switch, you configure the basic options needed to continue working with the switch, such as the IP address for the management interface, gateways, and some other options. Because the first boot is for a TOR switch is different from the one for the embedded version inside a BladeCenter chassis, we describe the differences and how to configure them in each case.

4.3 First boot of the RackSwitch G8264 switch

When using a TOR switch, before you first boot the switch, connect a console to it to be able to see the boot POST messages and log on to it. For a TOR model, you need to connect the console cable to a serial port of your computer and to the console port of the switch. This serial console is the only available method for you to connect to the switch in the first stage.

Use the following settings for your terminal client configuration (shown in Figure 4-3):

- Default baud rate: 9,600 bps
- Data bits: 8
- Stop bits: 1
- Parity: None
- Flow control: None

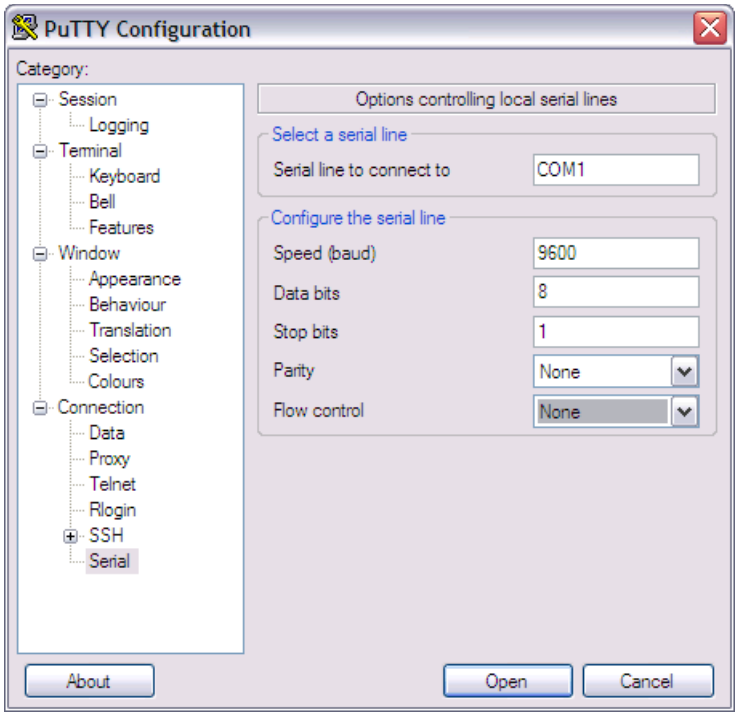

Figure 4-3 Serial terminal configuration

Typically, you need a serial to USB adapter, because most modern computers do not use serial ports. The switch-side connector is a mini-USB type connector. After connecting this cable to your system, you must identify the serial port you are using, and prepare console software to connect to it.

For our demonstration, we used a Windows XP client, and we used the PuTTY open source client as the serial console client.

If you are accessing a BladeCenter switch module, the task would be easier, because perform the initial configuration by using the AMM interfaces, either web or CLI. We explain the usage of these interfaces in 4.4, “First boot of the Virtual Fabric 10Gb Switch Module embedded switch” on page 142.

Terminal connection: Depending on your guest operating system, the terminal might look different, and the connection options for the serial port might also change. See your operating system help for troubleshooting any issues with the terminal connection to the switches.

After your terminal console is ready, you can plug in the power cables of the switch. The switch automatically powers on. The switch starts the boot process and you see messages on the console, as shown in Figure 4-4

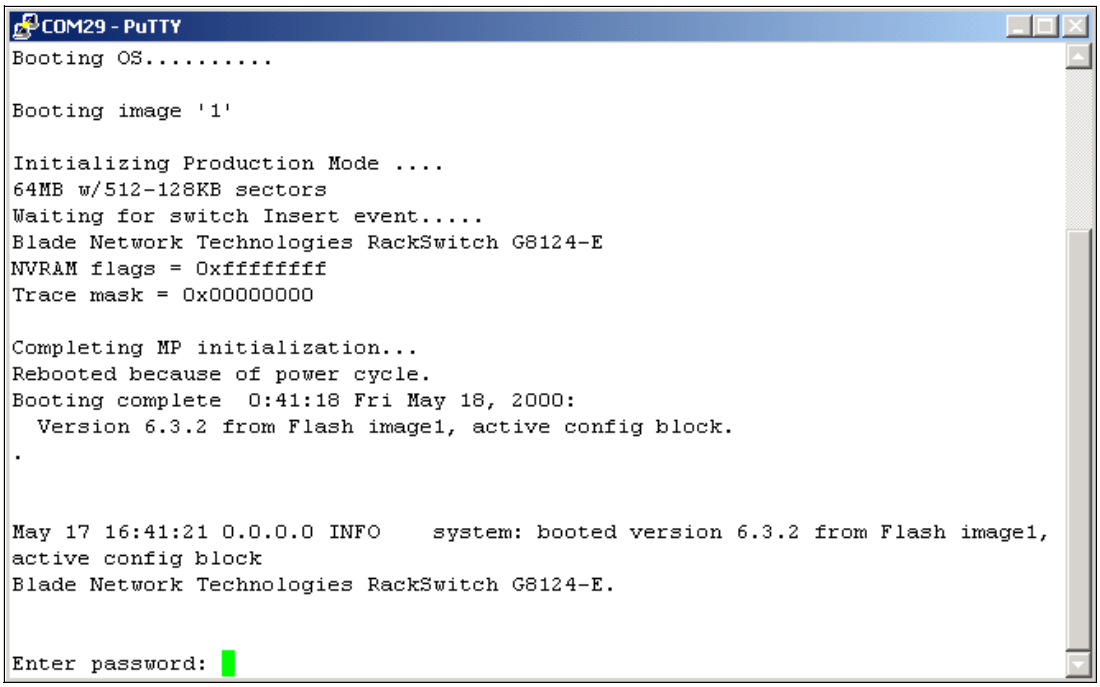
COM29 - PuTTY
Booting OS.....
Booting image '1'
Initializing Production Mode ....
64MB w/512-128KB sectors
Waiting for switch Insert event.....
Blade Network Technologies RackSwitch G8124-E
NVRAM flags = 0xffffffff
Trace mask = 0x00000000
Completing MP initialization...
Rebooted because of power cycle.
Booting complete 0:41:18 Fri May 18, 2000:
Version 6.3.2 from Flash image1, active config block.
May 17 16:41:21 0.0.0.0 INFO system: booted version 6.3.2 from Flash image1, active config block
Blade Network Technologies RackSwitch G8124-E.
Enter password:

Figure 4-4 Initial screen of an IBM RackSwitch G8124-E

4.3.1 Logging on to the switch

You must log on to the switch as admin to perform the initial setup.

Changing the password: The default password is “admin”. Remember to change this password to secure your infrastructure. For instructions about how to change the password, see “User management” on page 137.

If you connect by using SSH, the first time you connect to the switch, you must exchange encryption keys with the switch to establish the connection. Your SSH client handles this exchange automatically. In our case, the SSH client displayed a window to confirm the key exchange, as shown in Figure 4-5.

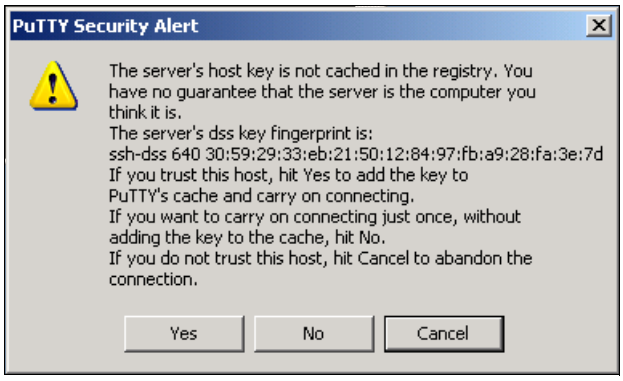
PuTTY Security Alert
The server's host key is not cached in the registry. You have no guarantee that the server is the computer you think it is.
The server's dss key fingerprint is:
ssh-dss 640 30:59:29:33:eb:21:50:12:84:97:fb:a9:28:fa:3e:7d
If you trust this host, hit Yes to add the key to PuTTY's cache and carry on connecting.
If you want to carry on connecting just once, without adding the key to the cache, hit No.
If you do not trust this host, hit Cancel to abandon the connection.
Yes
No
Cancel

Figure 4-5 SSH key exchange message

When you first log on to the switch, and if there no other user logged on to the switch, you are prompted to choose the CLI that you use. IBM switches support two kinds of CLI:

- IBM Networking OS, which is the default
- Industry standard CLI

You see a prompt similar to the one in Figure 4-6. Choose the CLI you want to use.

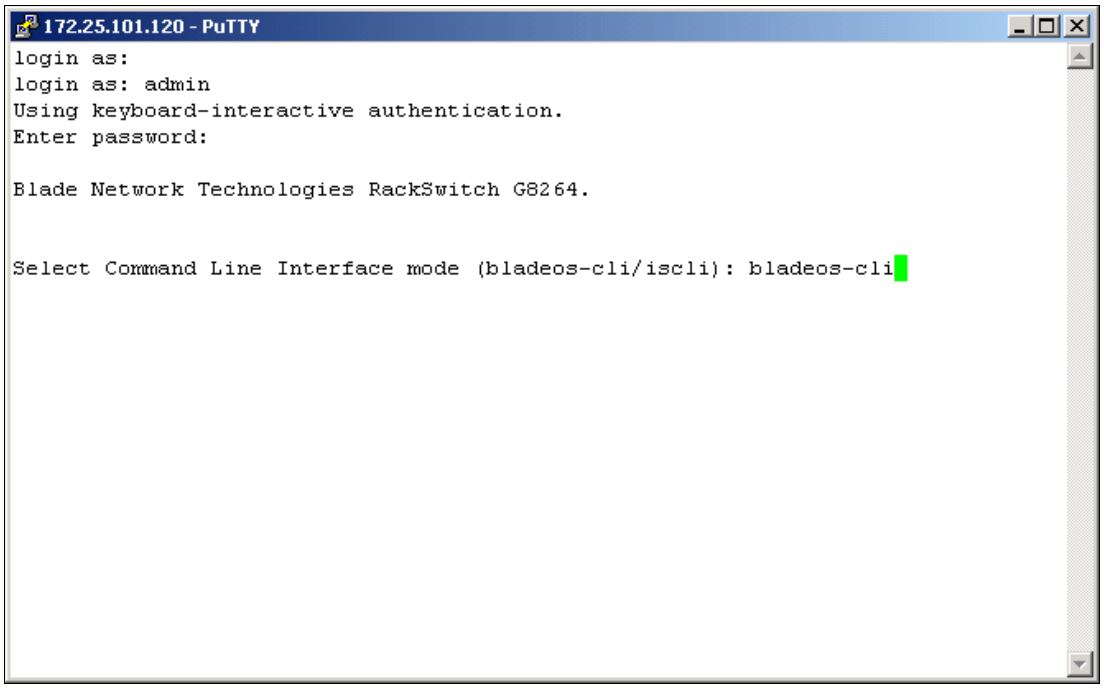

Figure 4-6 Initial prompt - CLI type selection

You then log on to the system. You can continue with the setup.

4.3.2 Global Configuration mode

In this scenario, we use the IBM Networking OS CLI and its commands. If we use the industry standard CLI, we note that we are using it where appropriate.

After you log on, enter the Global Configuration mode by running **enable** and then **configure terminal**, as shown in Example 4-1.

Example 4-1 Enabling the Global Configuration mode

```
RSG8264> enable

Enable privilege granted.
RSG8264# configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
RSG68264(config)#
```

Important: In the examples that we show in this section, we used the RackSwitch G8264, which is why the CLI shows RS8264 at the beginning of the prompt.

You can then configure the IP address of the management ports so you can access remote control with SSH or telnet. To accomplish this task, run the commands shown in Example 4-2. In our lab configuration, we used the interface ip 128 and we assigned the IPv4 address 172.25.101.120 with netmask 255.255.0.0.

Example 4-2 Configuring IP to management ports

```
RS8264(config)# interface ip [127|128]
RS8264(config-ip-if)# ip address <management interface IPv4 address>
RS8264(config-ip-if)# ip netmask <IPv4 subnet mask>
RS8264(config-ip-if)# enable
RS8264(config-ip-if)# exit
```

Configure the appropriate gateway (Example 4-3). In our lab configuration, the gateway is on address 172.25.1.1.

Example 4-3 Configuring gateway

```
RS8264(config)# ip gateway [3|4] address <default gateway IPv4 address>
RS8264(config)# ip gateway [3|4] enable
```

After configuring a management IP address for your switch, you can connect to a management port and use the Telnet program from an external management station to access and control the switch. The management port provides out-of-band management. You can also configure one of the other ports for in-band management. This process is explained in the *IBM G8264 Application Guide*, found at:

https://www-304.ibm.com/support/docview.wss?dc=DA400&rs=1126&uid=isg3T7000326&context=HW500&cs=utf-8&lang=en&loc=en_US

Connecting to the switch

After your switch is configured to have an IP address visible from your network, you can work with the switch remotely by using any Telnet or SSH client. Start a Telnet or SSH session and connect to the IP address that we defined in Example 4-2.

4.3.3 Setup tool

The IBM Networking OS includes a setup utility to complete the initial configuration of your switch. The setup utility prompts you to enter all the necessary information for the basic configuration of the switch. Whenever you log on as the system administrator under the factory default configuration, you are prompted whether you want to run the setup utility. Setup can also be activated manually from the CLI any time after you log on by running **/cfg/setup** (Example 4-4).

Example 4-4 Setup command interface

```
# /cfg/setup
```

Information needed for the setup utility

The setup utility requires the following information to perform the basic configuration:

- Basic system information:
  - Date & time
  - Whether to use Spanning Tree Group or not
- Optional configuration for each port:
  - Speed, duplex, flow control, and negotiation mode (as appropriate)
  - Whether to use VLAN tagging or not (as appropriate)
- Optional configuration for each VLAN:
  - Name of VLAN
  - Which ports are included in the VLAN
- Optional configuration of IP parameters:
  - IP address/mask and VLAN for each IP interface
  - IP addresses for default gateway
  - Whether IP forwarding is enabled or not

Important: If you do not enter a value when prompted, the setup utility uses the defaults defined by the IBM Networking OS, or leaves the requested value empty.

Basic configuration

The setup tool then prompts you with a series of questions. The first ones are the date and time, and whether to activate the Spanning Tree Protocol. In our case, we defined the date and time that corresponded to our timezone, and we defined Spanning Tree Group as `ON`. (Example 4-5).

Example 4-5 Date and time, and spanning tree configuration

```
"Set Up" will walk you through the configuration of
System Date and Time, Spanning Tree, Port Speed/Mode,
VLANs, and IP interfaces. [type Ctrl-C to abort "Set Up"]

System Date:
Enter year [2011]:

System Date:
Enter month [7]:

Enter day [22]:
```

```
System clock set to 18:55:36 Fri Jul 22, 2011.

System Time:
Enter hour in 24-hour format [18]:

Enter minutes [55]:

Enter seconds [37]:

System clock set to 18:55:36 Fri Jul 22, 2011.

Spanning Tree:
Current Spanning Tree Group 1 setting: ON
Turn Spanning Tree Group 1 OFF? [y/n]
```

The setup tool prompts you to configure VLANs and VLAN tagging for the ports (Example 4-6). If you want to change settings for VLANs, enter `y`, or enter `n` to skip VLAN configuration. In our case, we defined the VLANs used for our lab, as described in Chapter 3, “Reference architectures” on page 107.

Example 4-6 VLAN configuration

```
Port Config:
Will you configure VLANs and VLAN tagging for ports? [y/n]
```

After deciding if you want to configure the VLANs, you are prompted to configure the Gigabit Ethernet port flow parameters (Example 4-7).

Example 4-7 Gigabit Ethernet port flow parameters

```
Gig Link Configuration:
Port Flow Control:
Current Port EXT1 flow control setting: both
Enter new value ["rx"/"tx"/"both"/"none"]:
```

Enter `rx` to enable receive flow control, `tx` for transmit flow control, `both` to enable both, or `none` to turn off flow control for the port. To keep the current setting, press Enter.

If you select a port that has a Gigabit Ethernet connector, you can configure the port auto negotiation (Example 4-8).

Example 4-8 Port auto negotiation

```
Port Auto Negotiation:
Current Port EXT1 autonegotiation: on
Enter new value ["on"/"off"]:
```

Enter `on` to enable port autonegotiation, `off` to disable it, or press Enter to keep the current setting. In our lab configuration, we defined all the ports with the flow control disabled, with the option set to `none`, and the autonegotiation to `on`.

Depending on whether you answered yes or no to the VLAN configuration, you then define the different characteristics of the VLANS. If you selected to configure VLANs in Example 4-6 on page 132, you can enable or disable VLAN tagging for the port (Example 4-9).

Example 4-9 VLAN tagging configuration

```
Port VLAN tagging config (tagged port can be a member of multiple VLANs)
Current VLAN tag support: disabled
Enter new VLAN tag support [d/e]:
```

Enter `d` to disable VLAN tagging for the port or enter `e` to enable VLAN tagging for the port. To keep the current setting, press Enter.

The system prompts you to configure the next port (Example 4-10).

Example 4-10 Next port configuration

```
Enter port (INT1-14, MGT1-2, EXT1-64):
```

We use VLAN tagging on some of the ports. For more details about the precise configuration we used, see Chapter 3, “Reference architectures” on page 107.

The next steps are repeated for all the ports in the switch that you select in the prompt shown in Example 4-10. When you are done configuring the ports, press Enter without specifying any port.

If you want to change settings for individual VLANs, enter the number of the VLAN you want to configure (Example 4-11). To skip VLAN configuration, press Enter without entering a VLAN number.

Example 4-11 VLAN Config

```
VLAN Config:
Enter VLAN number from 2 to 4094, NULL at end:
```

Entering a new VLAN name is optional (Example 4-12). To use the pending new VLAN name, press Enter.

Example 4-12 VLAN name

```
Current VLAN name: VLAN 2
Enter new VLAN name:
```

Enter each port, by port number or port alias, and confirm the placement of the port into this VLAN (Example 4-13). When you are finished adding ports to this VLAN, press Enter without specifying any port.

Example 4-13 Define ports in a VLAN

```
Define Ports in VLAN:
Current VLAN 2: empty
Enter ports one per line, NULL at end:
```

After the VLANs are configured, you must configure the Spanning Tree Group membership for the VLAN. Follow the prompts from the setup tool shown in the Example 4-14.

Example 4-14 Spanning Tree membership

```
Spanning Tree Group membership:
Enter new Spanning Tree Group index [1-127]:
```

The system prompts you to configure the next VLAN (Example 4-15). If you want to configure another VLAN, enter the number.

Example 4-15 Next VLAN configuration

```
VLAN Config:
Enter VLAN number from 2 to 4094, NULL at end:
```

Repeat the steps in this section until all VLANs are configured. When all VLANs are configured, press Enter without specifying any VLAN to stop the VLAN configuration.

Configuration: All the details about the configuration of each port, VLANs, and features are described in Chapter 3, “Reference architectures” on page 107.

IP configuration

The setup tool now prompts for the IPv4 parameters. Although the switch supports both IPv4 and IPv6 networks, the setup utility permits only IPv4 configuration.

IP interfaces are used for defining the networks to which the switch belongs. Up to 128 IP interfaces can be configured on the RackSwitch G8264 (G8264). The IP address assigned to each IP interface provides the switch with an IP presence on your network. No two IP interfaces can be on the same IP network. The interfaces can be used for connecting to the switch for remote configuration, and for routing between subnets and VLANs (if used).

Important: IP interface 128 is reserved for out-of-band switch management.

Select the IP interface to configure, or skip interface configuration at the prompt (Example 4-16).

Example 4-16 Interface configuration

```
IP Config:
IP interfaces:
Enter interface number: (1-128)
```

In this example, the IP configured was `172.25.101.120` on interface `128`, which is the management port for external ports. You should configure each port you want to configure.

If you want to configure individual IP interfaces, enter the number of the IP interface you want to configure. To skip IP interface configuration, press Enter without typing an interface number.

For the specified IP interface, enter the IP address in IPv4 dotted decimal notation (Example 4-17).

Example 4-17 IP address configuration

```
Current IP address: 172.25.101.120
Enter new IP address:
```

To keep the current setting, press Enter.

At the prompt, enter the IPv4 subnet mask in dotted decimal notation (Example 4-18).

Example 4-18 Subnet mask configuration

```
Current subnet mask: 255.255.0.0
Enter new subnet mask:
```

To keep the current setting, press Enter. In our case, we used the `255.255.0.0` subnet.

If configuring VLANs, specify a VLAN for the interface (this prompt appears if you chose to configure VLANs) (Example 4-19).

Example 4-19 Specify VLAN

```
Current VLAN: 1
Enter new VLAN [1-4094]:
```

Enter the number for the VLAN to which the interface belongs, or press <Enter> without specifying a VLAN number to accept the current setting.

At the prompt, enter `y` to enable the IP interface, or `n` to leave it disabled (Example 4-20).

Example 4-20 Enable IP interface

```
Enable IP interface? [y/n]
```

The system prompts you to configure another interface (Example 4-21).

Example 4-21 Next interface configuration

```
Enter interface number: (1-128)
```

Repeat the steps in this section until all IP interfaces are configured. When all the interfaces are configured, press Enter without specifying any interface number.

Default gateway

At the prompt, select an IP default gateway for configuration, or skip default gateway configuration (Example 4-22).

Example 4-22 Default gateway

```
IP default gateways:
Enter default gateway number: (1-4)
```

Enter the number for the IP default gateway to be configured. In our lab setup, the gateway is `172.25.1.1`. To skip default gateway configuration, press <Enter> without typing a gateway number.

At the prompt, enter the IPv4 address for the selected default gateway (Example 4-23).

Example 4-23 Default gateway IP

```
Current IP address: 0.0.0.0
Enter new IP address:
```

Enter the IPv4 address in dotted decimal notation, or press <Enter> without specifying an address to accept the current setting.

At the prompt, enter `y` to enable the default gateway, or `n` to leave it disabled (Example 4-24).

Example 4-24 Enable default gateway

```
Enable default gateway? [y/n]
```

If you answer `yes`, the system prompts you to configure another default gateway (Example 4-25).

Example 4-25 Next default gateway configuration

```
Enter default gateway number: (1-4)
```

Repeat the steps in this section until all the default gateways are configured, in case you want to configure more than one. When all default gateways are configured, press <Enter> without specifying any number. In our lab setup, we specified only one default gateway.

IP routing

When IP interfaces are configured for the various IP subnets attached to your switch, IP routing between them can be performed entirely within the switch. This setup eliminates the need to send inter-subnet communication to an external router device. Routing on more complex networks, where subnets might not have a direct presence on the RackSwitch G8264, can be accomplished through configuring static routes or by letting the switch learn routes dynamically.

This part of the setup program prompts you to configure the various routing parameters. At the prompt, enable or disable forwarding for IP Routing (Example 4-26).

Example 4-26 Enable IP forwarding

```
Enable IP forwarding? [y/n]
```

Enter `y` to enable IP forwarding. To disable IP forwarding, enter `n`. To keep the current setting, press Enter.

Final steps

When prompted, decide whether to restart the setup or continue (Example 4-27).

Example 4-27 Restart Setup

```
Would you like to run from top again? [y/n]
```

Enter `y` to restart the Setup utility from the beginning, or `n` to continue.

When prompted, decide whether you want to review the configuration changes (Example 4-28).

Example 4-28 Review changes

```
Review the changes made? [y/n]
```

Enter `y` to review the changes made during this session of the setup utility. Enter `n` to continue without reviewing the changes.

Next, decide whether to apply the changes (Example 4-29).

Example 4-29 Apply changes

```
Apply the changes? [y/n]
```

Enter `y` to apply the changes, or `n` to continue without applying. Changes are normally applied.

At the prompt, decide whether to make the changes permanent (Example 4-30).

Example 4-30 Save changes to flash

```
Save changes to flash? [y/n]
```

Enter `y` to save the changes to flash. Enter `n` to continue without saving the changes. Changes are normally saved.

If you do not apply or save the changes, the system prompts you whether to abort them (Example 4-31).

Example 4-31 Abort changes

```
Abort all changes? [y/n]
```

Enter `y` to discard the changes. Enter `n` to return to the “Apply the changes?” prompt.

Important: After initial configuration is complete, change the default passwords.

In our case, we applied and saved all the changes, to have a permanent configuration that we could use for the lab environment and other tests in this book.

User management

IBM Networking OS allows an administrator to define user accounts that permit users to perform operation tasks through the switch by using CLI commands. After user accounts are configured and enabled, the switch requires user name and password authentication.

For example, an administrator can assign a user, who can then log on to the switch and perform operational commands (effective only until the next switch reboot). There is no need to have the administrator to perform all the maintenance tasks of the switch, or some of the basic configurations.

You should change the default passwords for your administrator users. To accomplish this task, you must perform the commands shown in Example 4-32.

Example 4-32 Change password procedure for admin user

```
>> RS8264 - Main# /cfg/sys/access
>> RS8264 - System Access# user/admpw
Changing ADMINISTRATOR password; validation required:
Enter current admin password:
Enter new admin password (max 128 characters):
Re-enter new admin password:
New admin password accepted.

>> RS8264 - System Access# apply
>> RS8264 - System Access# save
```

Remember to apply and save your changes so the password is changed in the system.

If you want to create a user, run the commands shown in Example 4-33. You define the access rights to the user you want to create, and then provide a password for it. You are prompted for the admin password to validate the creation.

Example 4-33 Adding a user

```
RS8264(config)# access user 1 name <1-8 characters>
RS8264(config)# access user 1 password
Changing user1 password; validation required:
Enter current admin password: <current administrator password>
Enter new user1 password: <new user password>
Re-enter new user1 password: <new user password>
New user1 password accepted.
```

In our lab setup, we kept the default users, which were enough for us to do all the tasks required for this book. You might have other needs or requirements that require you to create some more users.

The user is by default assigned to the user access level (also known as class of service (COS)). COS for all user accounts have global access to all resources except for User COS, which has access to view only resources that the user owns. To change the user's level, select one of the options shown in Example 4-34.

Example 4-34 User access level

```
RS8264(config)# access user 1 level {user|operator|administrator}
```

The user levels are defined in the Table 4-1.

Table 4-1 User types

User Account	Description and tasks performed	Default password
User	The User has no direct responsibility for switch management. The User can view all switch status information and statistics, but cannot make any configuration changes to the switch.	user

User Account	Description and tasks performed	Default password
Operator	The Operator manages all functions of the switch. The Operator can reset ports, except for the management port.	operator
Administrator	The super-user Administrator has complete access to all commands, information, and configuration commands on the switch, including the ability to change both the user and administrator passwords.	admin

To confirm that the creation of the user is done correctly, run **show access** (Example 4-35).

Example 4-35 Adding a user

```
RS8264# show access user uid 1
```

You can also enable or disable one user (Example 4-36). A user account must be enabled before the switch recognizes and permits login under the account. After it is enabled, the switch requires any user to enter both a user name and password.

Example 4-36 Enable or disable a user

```
RS8264(config)# [no] access user 1 enable
```

You can list the existing users in the switch by running **show access user** (Example 4-37).

Example 4-37 Showing the users of the system

```
RS8264# show access user
Usernames:
user - Enabled - offline
oper - Disabled - offline
admin - Always Enabled - online 1 session
Current User ID table:
1: name jane , ena, cos user , password valid, online 1 session
2: name john , ena, cos user , password valid, online 2 sessions
```

IBM Networking OS also supports other authentication systems, such as RADIUS, TACACS+, and LDAP. With these security solutions, the user management is done external to the switch, so the only thing that must be done is to configure the switch to access the external security server.

RADIUS

Use the following procedure to configure RADIUS authentication on your switch.

Configure the IPv4 addresses of the Primary and Secondary RADIUS servers, and enable RADIUS authentication (Example 4-38). In our case, the RADIUS server is in the IP address 10.10.1.1 and the secondary host in the 10.10.1.2.

Example 4-38 RADIUS authentication

```
RS8264(config)# radius-server primary-host 10.10.1.1
RS8264(config)# radius-server secondary-host 10.10.1.2
RS8264(config)# radius-server enable
```

You must configure the RADIUS secret by running the commands shown in Example 4-39.

Example 4-39 RADIUS secret

```
RS8264(config)# radius-server primary-host 10.10.1.1 key <1-32 character secret>
RS8264(config)# radius-server secondary-host 10.10.1.2 key <1-32 character secret>
```

You may change the default UDP port number used to listen to RADIUS (Example 4-40). The known port for RADIUS is 1812.

Example 4-40 RADIUS UDP port

```
RS8264(config)# radius-server port <UDP port number>
```

Configure the number of retry attempts for contacting the RADIUS server, and the timeout period (Example 4-41).

Example 4-41 RADIUS retry and timeout

```
RS8264(config)# radius-server retransmit 3
RS8264(config)# radius-server timeout 5
```

RADIUS options: For more detailed information about all the options related to RADIUS, see IBM RackSwitch G8264 Blade OS Application Guide, found at:

http://www-01.ibm.com/support/docview.wss?uid=isg3T7000326

TACACS+

When TACACS+ Command Authorization is enabled, IBM Networking OS configuration commands are sent to the TACACS+ server for authorization. Use the command shown in Example 4-42 to enable TACACS+ Command Authorization.

Example 4-42 TACACS+ Command Authorization

```
RS8264(config)# tacacs-server command-authorization
```

When TACACS+ Command Logging is enabled, IBM Networking OS configuration commands are logged on the TACACS+ server. Use the command shown in Example 4-43 to enable TACACS+ Command Logging.

Example 4-43 TACACS+ Command Logging

```
RS8264(config)# tacacs-server command-logging
```

The examples shown in Example 4-44 illustrate the format of IBM Networking OS commands sent to the TACACS+ server.

Example 4-44 Command format

```
authorization request, cmd=shell, cmd-arg=interface ip
accounting request, cmd=shell, cmd-arg=interface ip
authorization request, cmd=shell, cmd-arg=enable
accounting request, cmd=shell, cmd-arg=enable
```

As shown in Example 4-45, configure the IPv4 addresses of the Primary and Secondary TACACS+ servers, and enable TACACS authentication. Specify the interface port (optional).

Example 4-45 Primary and Secondary TACACS+ servers

```
RS8264(config)# tacacs-server primary-host 10.10.1.1
RS8264(config)# tacacs-server primary-host mgt-port
RS8264(config)# tacacs-server secondary-host 10.10.1.2
RS8264(config)# tacacs-server secondary-host data-port
RS8264(config)# tacacs-server enable
```

Configure the TACACS+ secret and second secret (Example 4-46). In our example, the primary host is in IP 10.10.1.1 and the secondary in 10.10.1.2.

Example 4-46 TACACS+ secret

```
RS8264(config)# tacacs-server primary-host 10.10.1.1 key <1-32 character secret>
RS8264(config)# tacacs-server secondary-host 10.10.1.2 key <1-32 character secret>
```

You may change the default TCP port number used to listen to TACACS+ (Example 4-47). The known port for TACACS+ is 49.

Example 4-47 TACACS+ TCP port

```
RS8264(config)# tacacs-server port <TCP port number>
```

Configure the number of retry attempts, and the timeout period (Example 4-48).

Example 4-48 TACACS+ retry and timeout

```
RS8264(config)# tacacs-server retransmit 3
RS8264(config)# tacacs-server timeout 5
```

LDAP

To configure the LDAP access, complete the steps in this section.

As shown in Example 4-49, turn LDAP authentication on, then configure the IPv4 addresses of the Primary and SecondaryLDAP servers. Specify the interface port (optional). In our example, the primary host is in IP 10.10.1.1 and the secondary in 10.10.1.2.

Example 4-49 LDAP configuration

```
>> # ldap-server enable
>> # ldap-server primary-host 10.10.1.1 mgt-port
>> # ldap-server secondary-host 10.10.1.2 data-port
```

Configure the domain name (Example 4-50).

Example 4-50 Domain name

```
>> # ldap-server domain <ou=people,dc=my-domain,dc=com>
```

You may change the default TCP port number used to listen to LDAP (optional) (Example 4-51). The known port for LDAP is 389.

Example 4-51 LDAP port

```
>> # ldap-server port <1-65000>
```

Configure the number of retry attempts for contacting the LDAP server, and the timeout period (Example 4-52).

Example 4-52 LDAP retry and timeout

```
>> # ldap-server retransmit 3
>> # ldap-server timeout 10
```

4.4 First boot of the Virtual Fabric 10Gb Switch Module embedded switch

If the switch is the embedded model inside the BladeCenter chassis, the initial boot and setup is different. The management interfaces are accessible from the Advanced Management Module (AMM) of the BladeCenter chassis. You should first log on to the AMM and then configure the I/O modules that correspond to the switches you want to configure. During this section, we describe this process and use the web interface. The CLI is also available from a Telnet session, but using it is similar to the procedure described in 4.3, “First boot of the RackSwitch G8264 switch” on page 127.

4.4.1 Basic options

Log on to the AMM by using your user name and password. Once in the web interface, expand **I/O Module tasks** and click **Admin/Power/Restart**. The window shown in Figure 4-7 opens. Here you can configure a fast POST process and enable the external ports, which accelerate the boot time of the switch and provides access to the external ports of it. Use the settings shown in the figure.

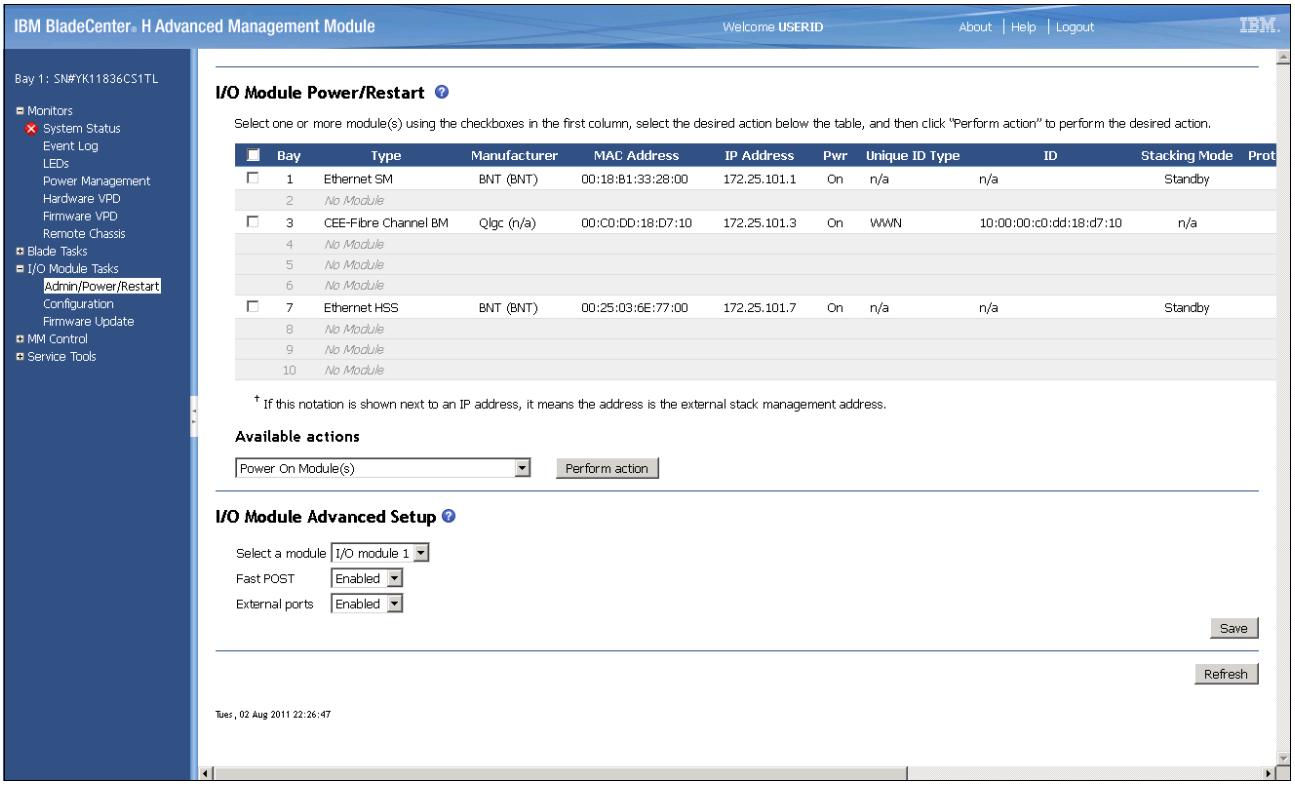
IBM BladeCenter H Advanced Management Module
Welcome USERID
About | Help | Logout
IBM.
Bay 1: SN#YK11836CS1TL
Monitors
System Status
Event Log
LEDs
Power Management
Hardware VPD
Firmware VPD
Remote Chassis
Blade Tasks
I/O Module Tasks
Admin/Power/Restart
Configuration
Firmware Update
MM Control
Service Tools
I/O Module Power/Restart
Select one or more module(s) using the checkboxes in the first column, select the desired action below the table, and then click "Perform action" to perform the desired action.
Bay | Type | Manufacturer | MAC Address | IP Address | Pwr | Unique ID Type | ID | Stacking Mode | Prot
1 | Ethernet SM | BNT (BNT) | 00:18:B1:33:28:00 | 172.25.101.1 | On | n/a | n/a | Standby
2 | No Module
3 | CEE-Fibre Channel BM | Qlgc (n/a) | 00:C0:DD:18:D7:10 | 172.25.101.3 | On | WWN | 10:00:00:c0:dd:18:d7:10 | n/a
4 | No Module
5 | No Module
6 | No Module
7 | Ethernet HSS | BNT (BNT) | 00:25:03:6E:77:00 | 172.25.101.7 | On | n/a | n/a | Standby
8 | No Module
9 | No Module
10 | No Module
† If this notation is shown next to an IP address, it means the address is the external stack management address.
Available actions
Power On Module(s)
Perform action
I/O Module Advanced Setup
Select a module
I/O module 1
Fast POST
Enabled
External ports
Enabled
Save
Refresh
Tues, 02 Aug 2011 22:26:47

Figure 4-7 AMM web interface

After you complete the POST process and enable the ports, click **Configuration**, and the window shown in Figure 4-8 opens.

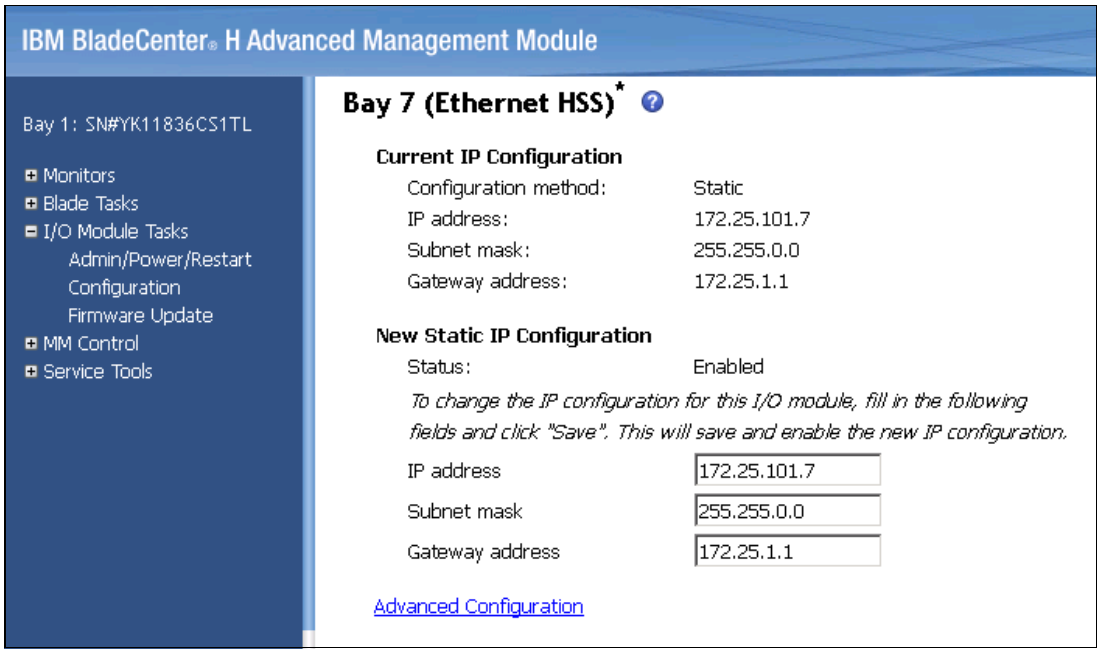

Figure 4-8 I/O Module tasks menu on the AMM

From this menu, you can identify the switch on one of the high speed bays. In our case, the High Speed Switch (HSS) was installed in bay 7 of the BladeCenter-H chassis.

4.4.2 Setting an IP address

In Figure 4-8, you can also configure the management IP address, the subnet mask, and the gateway to make the switch visible to the management network. This IP address is the one that you connect to over telnet or SSH to access the console. Use the settings shown in Figure 4-8.

4.4.3 Advanced options

If you click Advanced Configuration in Figure 4-8 on page 144, the window shown in Figure 4-9 opens.

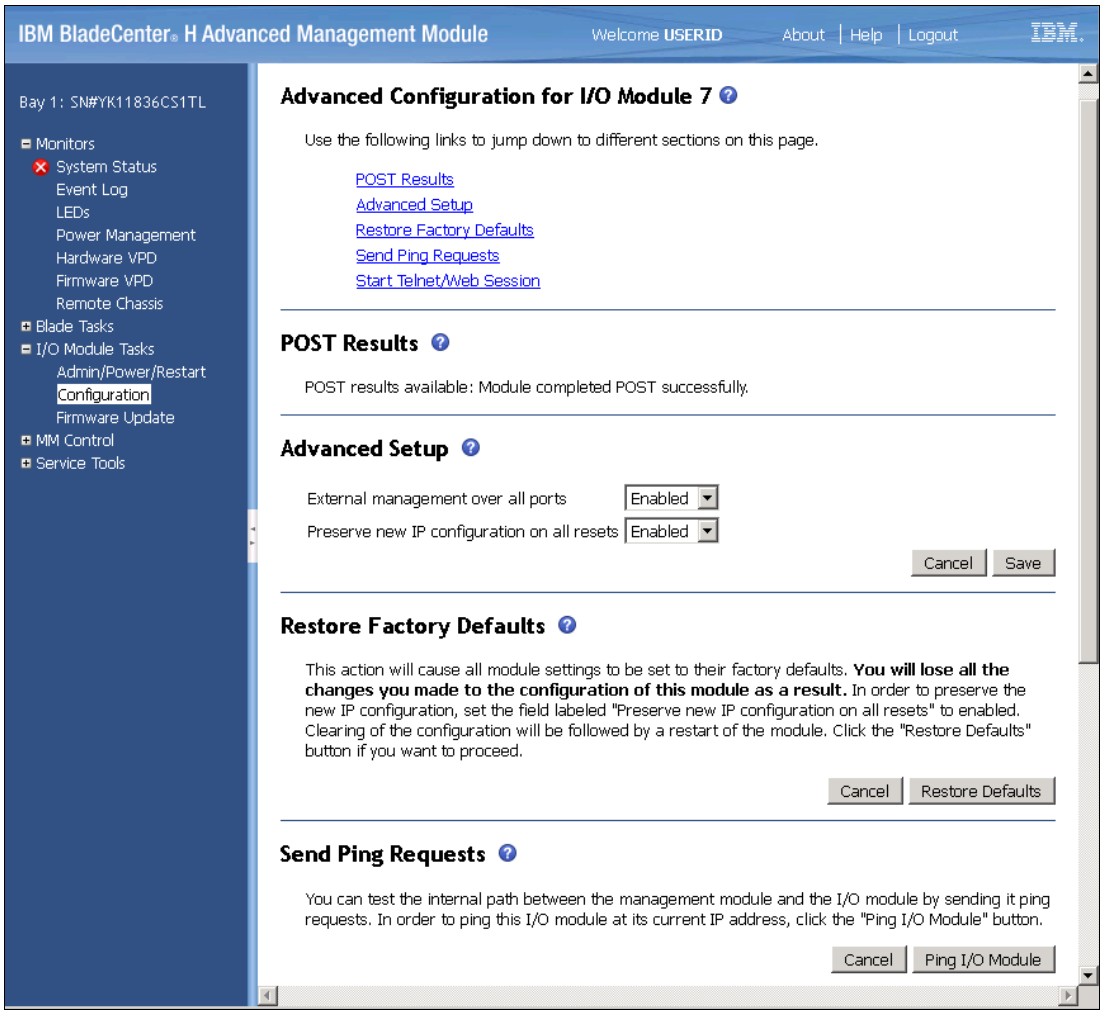
IBM BladeCenter H Advanced Management Module
Welcome USERID
About | Help | Logout
IBM.
Bay 1: SN#YK11836CS1TL
Monitors
System Status
Event Log
LEDs
Power Management
Hardware VPD
Firmware VPD
Remote Chassis
Blade Tasks
I/O Module Tasks
Admin/Power/Restart
Configuration
Firmware Update
MM Control
Service Tools
Advanced Configuration for I/O Module 7
Use the following links to jump down to different sections on this page.
POST Results
Advanced Setup
Restore Factory Defaults
Send Ping Requests
Start Telnet/Web Session
POST Results
POST results available: Module completed POST successfully.
Advanced Setup
External management over all ports
Enabled
Preserve new IP configuration on all resets
Enabled
Cancel
Save
Restore Factory Defaults
This action will cause all module settings to be set to their factory defaults. You will lose all the changes you made to the configuration of this module as a result. In order to preserve the new IP configuration, set the field labeled "Preserve new IP configuration on all resets" to enabled. Clearing of the configuration will be followed by a restart of the module. Click the "Restore Defaults" button if you want to proceed.
Cancel
Restore Defaults
Send Ping Requests
You can test the internal path between the management module and the I/O module by sending it ping requests. In order to ping this I/O module at its current IP address, click the "Ping I/O Module" button.
Cancel
Ping I/O Module

Figure 4-9 Advanced I/O modules configuration window

Here you can perform more advanced configuration and access more detailed information about the switch, such as:

- POST results (to view the start messages)
- Advanced setup (to enable advanced features)
- Restore factory defaults
- Send **`ping`** requests (to check that the switch receives **`ping`** commands)
- Start a Telnet/web session (to manage the switch by using IBM Networking OS), which is shown in Figure 4-10 on page 146.

In this book, we focus only on the options that are relevant to our topic. Each of these options is documented in the BladeCenter manuals and in *IBM BladeCenter Products and Technology*, SG24-7523.

In the Advanced Configuration window, you can define the management ports to be visible from the external ports. Depending on your needs, this action might be necessary to access the switches. Typically, you enable this feature, unless you want to handle the switches only from the AMM interface. We do not describe how to perform this procedure in this book.

4.4.4 Telnet access

You can start a telnet session from the window shown in Figure 4-9 on page 145. This session is similar to any TOR switch. To learn how to enable Telnet access, see “Connecting to the switch” on page 130.

The telnet session is shown in Figure 4-10.

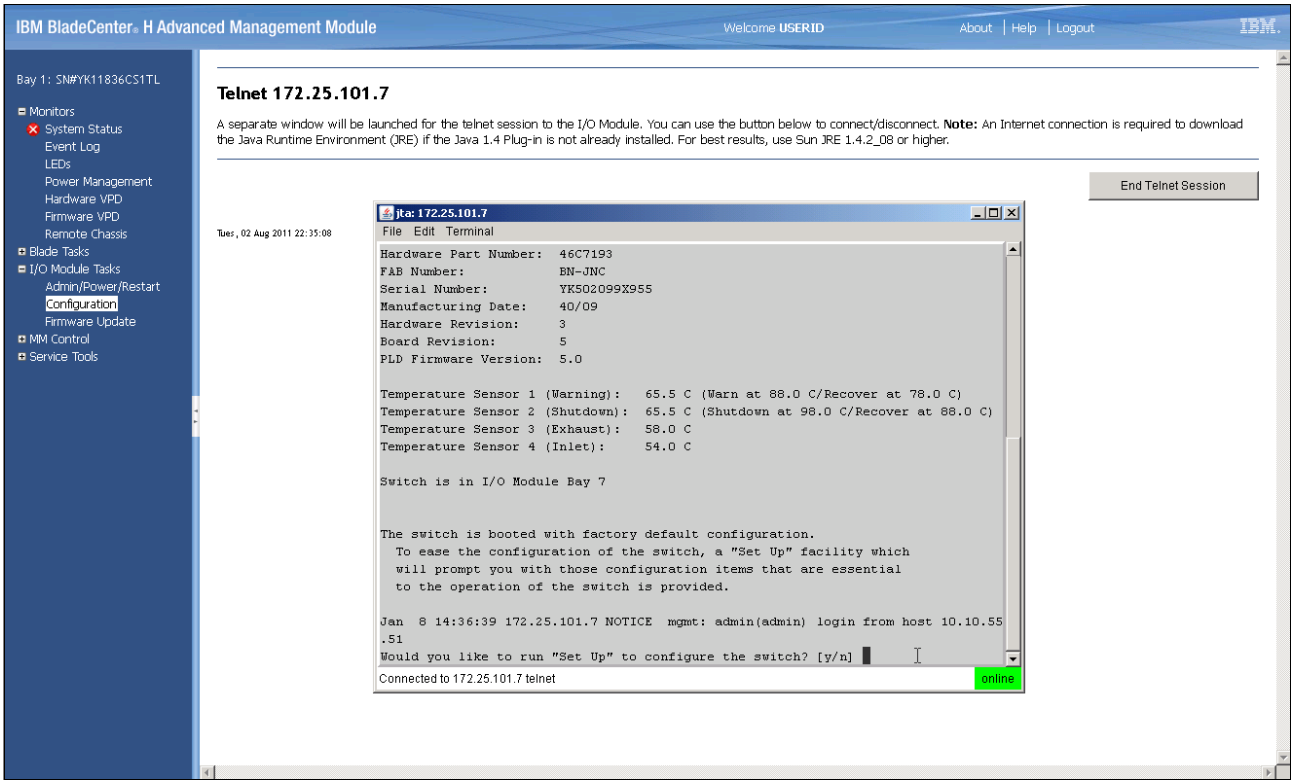
IBM BladeCenter H Advanced Management Module
Welcome USERID
About | Help | Logout
IBM.
Bay 1: SN#YK11836CS1TL
Monitors
System Status
Event Log
LEDs
Power Management
Hardware VPD
Firmware VPD
Remote Chassis
Blade Tasks
I/O Module Tasks
Admin/Power/Restart
Configuration
Firmware Update
MM Control
Service Tools
Telnet 172.25.101.7
A separate window will be launched for the telnet session to the I/O Module. You can use the button below to connect/disconnect. Note: An Internet connection is required to download the Java Runtime Environment (JRE) if the Java 1.4 Plug-in is not already installed. For best results, use Sun JRE 1.4.2_08 or higher.
End Telnet Session
Tues, 02 Aug 2011 22:35:08
jta: 172.25.101.7
File Edit Terminal
Hardware Part Number: 46C7193
FAB Number: BN-JNC
Serial Number: YK502099X955
Manufacturing Date: 40/09
Hardware Revision: 3
Board Revision: 5
PLD Firmware Version: 5.0
Temperature Sensor 1 (Warning): 65.5 C (Warn at 88.0 C/Recover at 78.0 C)
Temperature Sensor 2 (Shutdown): 65.5 C (Shutdown at 98.0 C/Recover at 88.0 C)
Temperature Sensor 3 (Exhaust): 58.0 C
Temperature Sensor 4 (Inlet): 54.0 C
Switch is in I/O Module Bay 7
The switch is booted with factory default configuration.
To ease the configuration of the switch, a "Set Up" facility which
will prompt you with those configuration items that are essential
to the operation of the switch is provided.
Jan 8 14:36:39 172.25.101.7 NOTICE mgmt: admin(admin) login from host 10.10.55
.51
Would you like to run "Set Up" to configure the switch? [y/n]
Connected to 172.25.101.7 telnet
online

Figure 4-10 Telnet session to the switch from the AMM

To end the session, just type **`exit`** and close the window.

4.4.5 Web access

You can access the web interface of the switch directly from the window shown in Figure 4-9 on page 145. Click **Start Web Session**, and a separate window on your web browser opens. From this window, you log in with your user name and password (Figure 4-11).

Figure 4-11 Web interface web login window

After a successful login, you see the IBM Networking OS web interface dashboard (Figure 4-12).

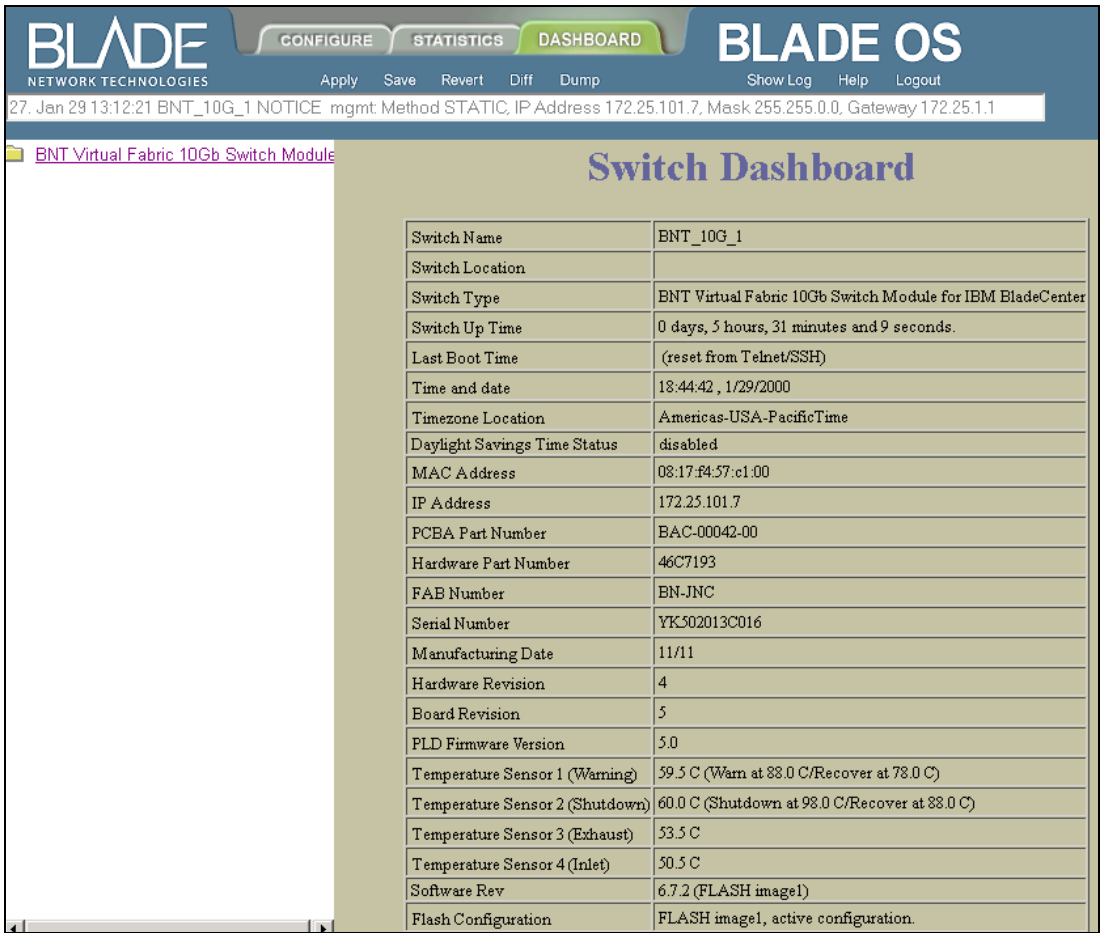
BLADE
NETWORK TECHNOLOGIES
CONFIGURE
STATISTICS
DASHBOARD
BLADE OS
Apply
Save
Revert
Diff
Dump
Show Log
Help
Logout
27. Jan 29 13:12:21 BNT_10G_1 NOTICE mgmt: Method STATIC, IP Address 172.25.101.7, Mask 255.255.0.0, Gateway 172.25.1.1
BNT Virtual Fabric 10Gb Switch Module
Switch Dashboard
Switch Name BNT_10G_1
Switch Location
Switch Type BNT Virtual Fabric 10Gb Switch Module for IBM BladeCenter
Switch Up Time 0 days, 5 hours, 31 minutes and 9 seconds.
Last Boot Time (reset from Telnet/SSH)
Time and date 18:44:42 , 1/29/2000
Timezone Location Americas-USA-PacificTime
Daylight Savings Time Status disabled
MAC Address 08:17:f4:57:c1:00
IP Address 172.25.101.7
PCBA Part Number BAC-00042-00
Hardware Part Number 46C7193
FAB Number BN-JNC
Serial Number YK502013C016
Manufacturing Date 11/11
Hardware Revision 4
Board Revision 5
PLD Firmware Version 5.0
Temperature Sensor 1 (Warning) 59.5 C (Warn at 88.0 C/Recover at 78.0 C)
Temperature Sensor 2 (Shutdown) 60.0 C (Shutdown at 98.0 C/Recover at 88.0 C)
Temperature Sensor 3 (Exhaust) 53.5 C
Temperature Sensor 4 (Inlet) 50.5 C
Software Rev 6.7.2 (FLASH image1)
Flash Configuration FLASH image1, active configuration.

Figure 4-12 IBM Networking OS web interface dashboard

From this interface, you can perform different configuration tasks on the switch. For more details and configuration options, see the appropriate documentation:

- *IBM RackSwitch G8264 Application Guide*:

  http://www-01.ibm.com/support/docview.wss?rs=1126&context=HW500&dc=DA400&uid=isg3T7000326&loc=en_US&cs=utf-8&lang=en

- *IBM Virtual Fabric 10Gb Switch Module for IBM BladeCenter - Installation Guide*:

  http://publib.boulder.ibm.com/infocenter/bladectr/documentation/topic/com.ibm.bladecenter.io_39Y9267.doc/81y1115.pdf

Important: Remember to submit, apply, and save your configuration changes.

4.4.6 Setting the date and time

The embedded switch does not synchronize the date and time with the BladeCenter chassis, so you must set it manually during the first boot. From the web session, choose your switch from the list of switches and click **System**. A window with a set of options opens (Figure 4-13). Set the date and time here.

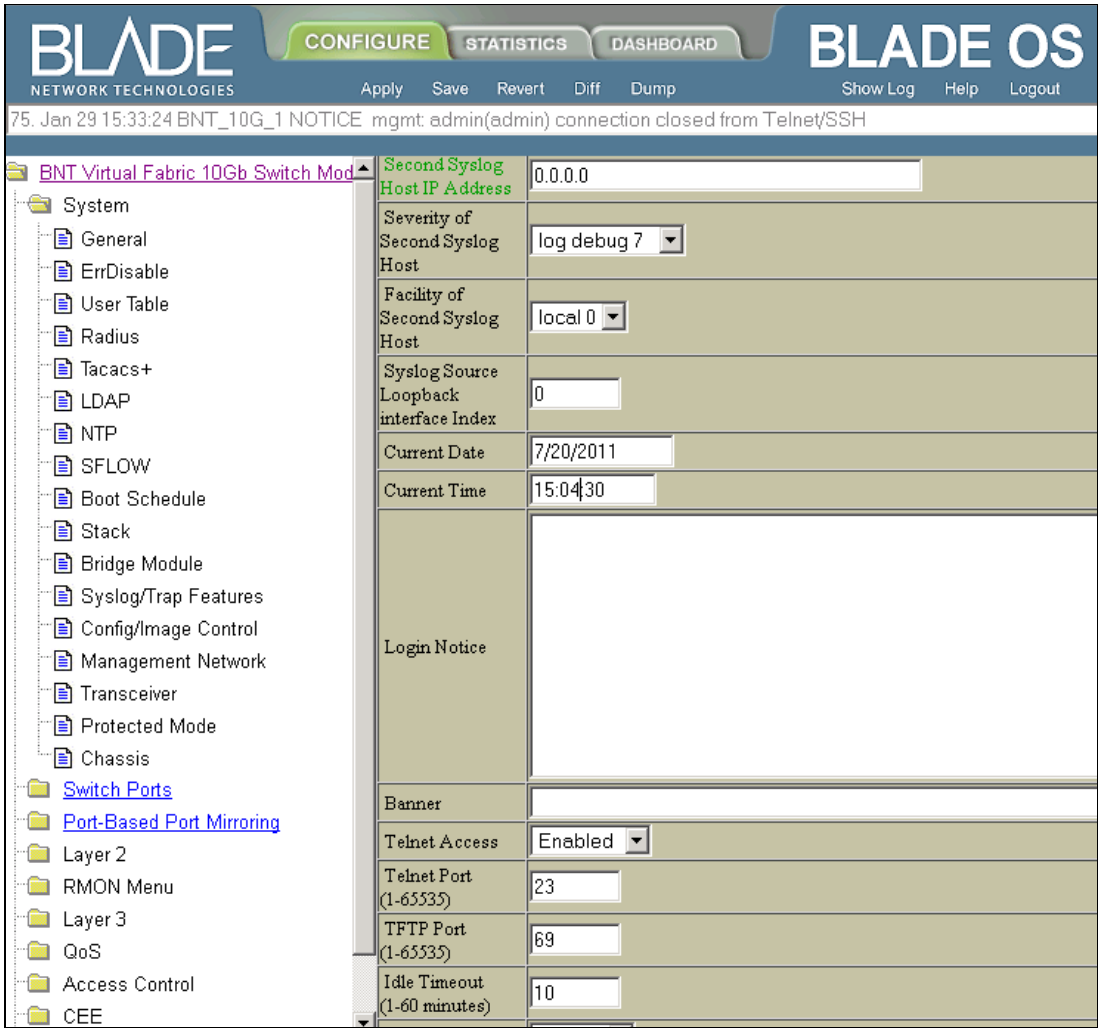

Figure 4-13 System configuration menu from the web interface

4.4.7 Firmware upgrade from the AMM web interface

The I/O Module Firmware Update option of the AMM web interface is currently not supported for the IBM 10Gb Switches. If you try to update the firmware using this option, you receive the message shown in Figure 4-14.

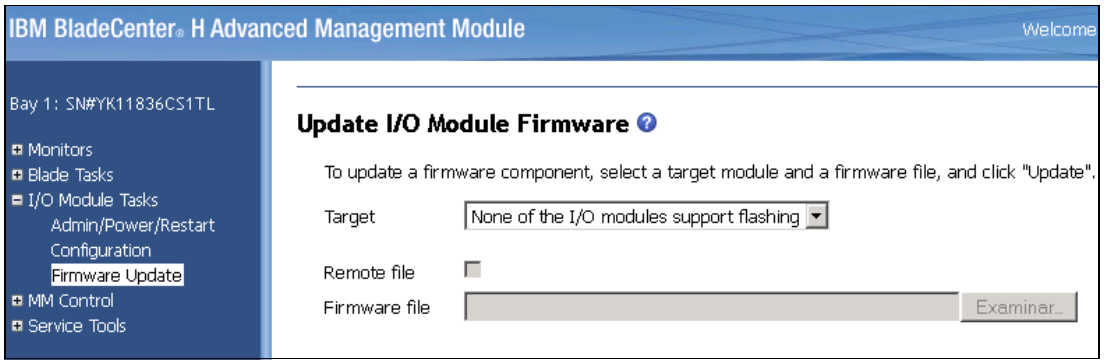

Figure 4-14 Firmware upgrade from the AMM web interface is not possible

4.4.8 Working with users and passwords

You can see users and passwords from the web interface. To do so, select the switch you want to configure on the left side, and click **System** → **User Table** to see the user configuration window, shown in Figure 4-15.

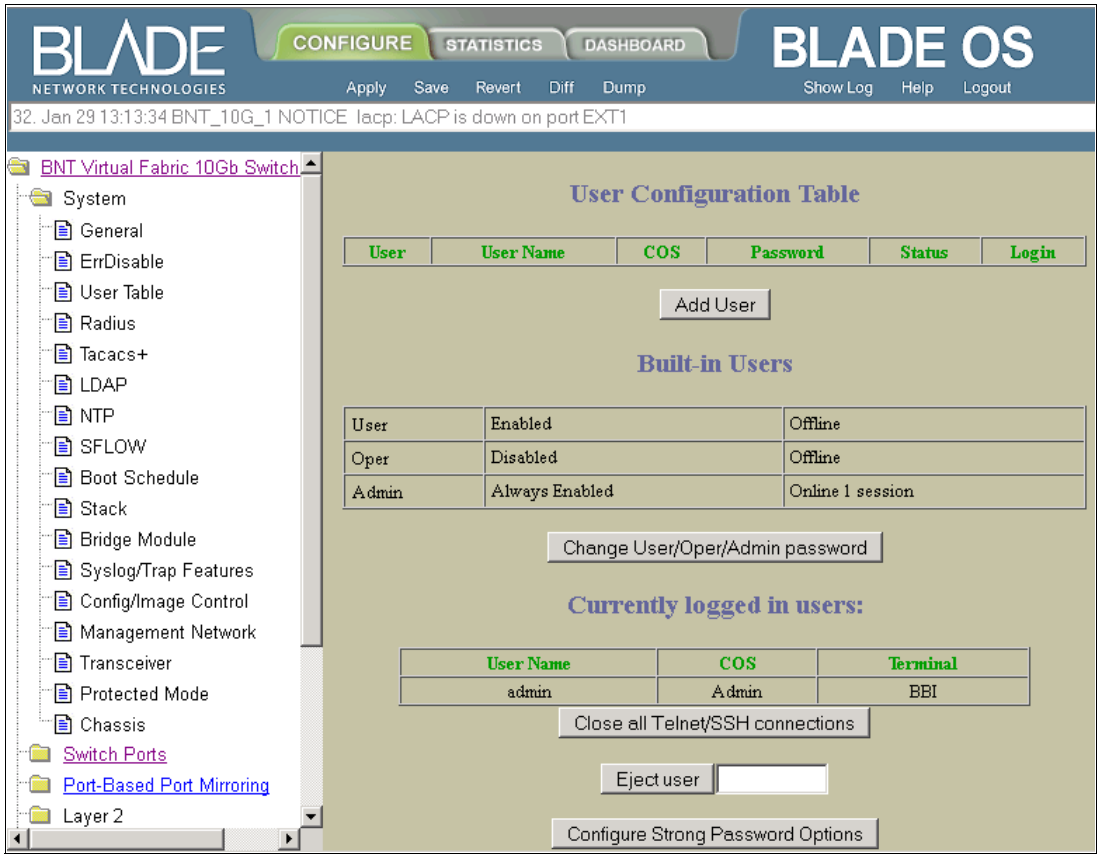

Figure 4-15 User configuration window from the web interface

Embedded switch specifics

In a TOR switch, you can add a user by clicking the **Add User** button. You cannot perform this function from the BladeCenter embedded switch module. If you try to do it, you receive an error message, because in the BladeCenter chassis, user creation is handled by the AMM, and you cannot use the web interface.

You can change passwords and work with users from the CLI, as described in “User management” on page 137.

4.5 IBM System Networking Element Manager

IBM System Networking Element Manager (SNEM) is an application for remote monitoring and management of Ethernet switches from IBM. It is designed to simplify and centralize the management of your BladeCenter or blade server and Top-of-Rack Ethernet switches. SNEM offers data center network managers a competitive solution to ensure availability of their IBM network devices used with business critical applications and services.

The increased scale of data centers leads to a larger number of physical and virtual devices in the data center. SNEM can help alleviate issues that result from this increase of devices. SNEM offers:

- Simplified management: Point and click administration of large groups of switches from a central location in a single operation. Updates that took hours or days can be done in minutes.
- Automation: Automated and scheduled software downloads, configuration updates/backups, switch reboots, and VM network policy distribution simplifies switch management and reduces human errors.
- Integration: Integration with industry-leading enterprise management systems (HP System Insight Manager (SIM) and IBM Systems Director), authentication servers, and virtualization management applications improve security and help reduce training and other costs.
- Monitoring and reporting: SNEM can monitor the health and performance of the network, and get comprehensive asset, event, and virtualization reporting on all IBM switches in the network.

4.5.1 IBM System Networking Element Manager solution architecture

The SNEM solution is a package that contains the following element and network management software that supports IBM networking hardware devices:

- IBM System Networking Element Manager (SNEM) V6.1
- IBM Tivoli® Network Manager V1.0 for System Networking Element Manager
- IBM Tivoli Netcool® Configuration Manager V6.3
- IBM Tivoli Netcool/OMNIbus V7.3.1

The various software components of SNEM are shown in Figure 4-16.

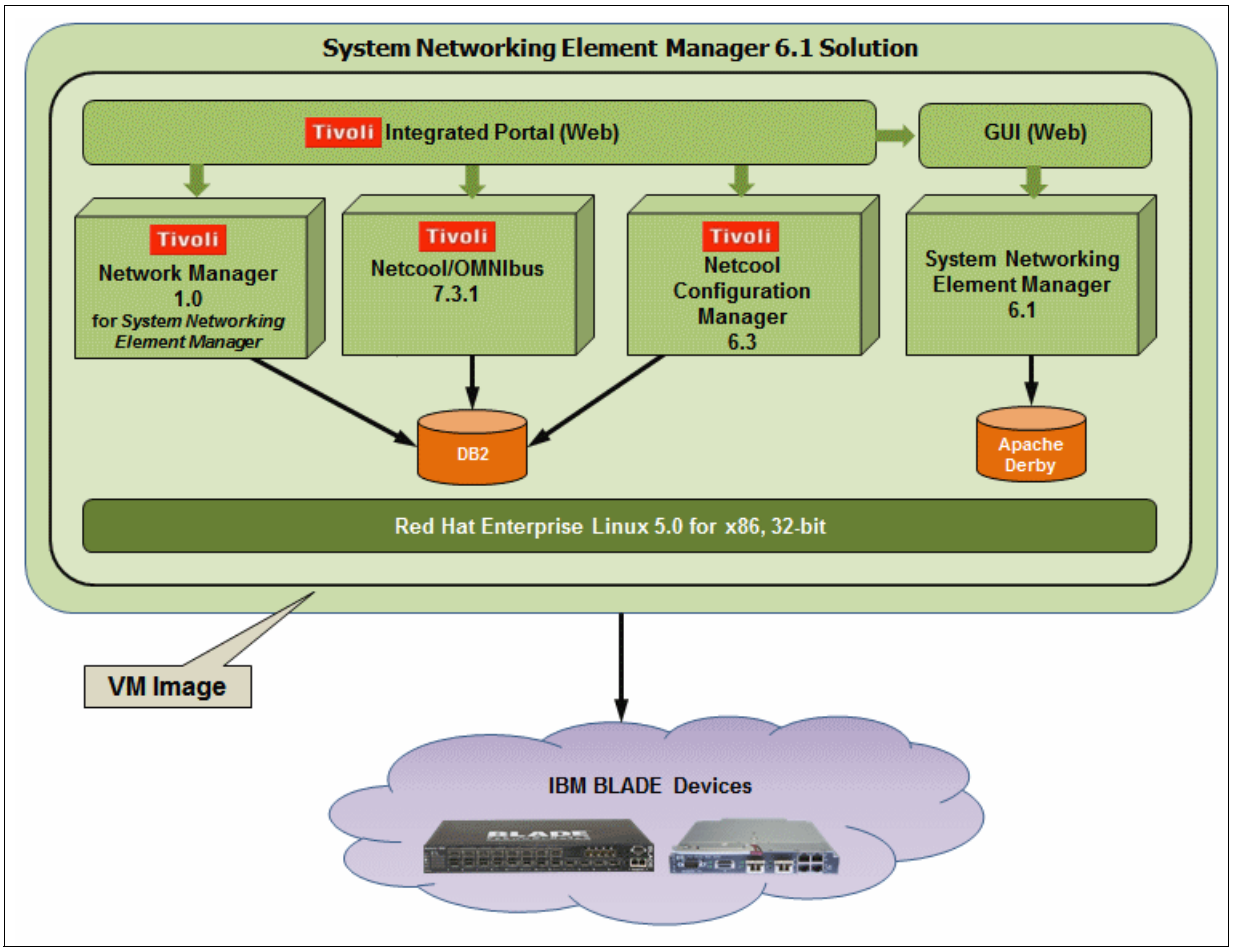

Figure 4-16 SNEM solution architecture

IBM System Networking Element Manager

SNEM provides a single point for management that allows automation of basic network tasks, including remote monitoring and management of Ethernet switches.

The benefits of SNEM include:

- Improves network visibility and drive reliability and performance.
- Increases the availability and performance of critical business services with advanced correlation, event deduplication, automated diagnostic tests, and root-cause analysis.
- Simplifies management of large groups of switches with automatic discovery of switches on the network.
- Automates and integrates management, deployment, and monitoring functions.

Tivoli Network Manager V1 for SNEM

Tivoli Network Manager provides the features necessary to manage complex networks. These features include network discovery, device polling, including storage of polled SNMP and ICMP data for reporting and analysis, and topology visualization.

Tivoli Network Manager can display network events, perform root-cause analysis of network events, and enrich network events with topology and other network data. Tivoli Network Manager integrates with other IBM products, such as IBM Tivoli Business Service Manager, Tivoli Application Dependency Discovery Manager, and IBM Systems Director.

Using Tivoli Network Manager, you can perform the following tasks:

- Manage complex networks.
- View the network in multiple ways.
- Apply ready-to-use device and interface polling capabilities.
- Use built-in root-cause analysis capabilities.
- Troubleshoot network problems using right-click tools.
- Generate richer network visualization and event data.
- Discover increasingly bigger networks.
- Run reports to retrieve essential network data.
- Build custom multi-portlet pages.

Tivoli Netcool Configuration Manager

Tivoli Netcool Configuration Manager provides configuration management support for network devices, including extensive configuration policy thresholding capabilities.

Tivoli Netcool/OMNIbus

The Tivoli Netcool/OMNIbus software collects and manages network event information, and delivers real-time, centralized monitoring of complex networks and IT domains.

Tivoli Netcool/OMNIbus tracks alert information in a high-performance, in-memory database, and presents information of interest to specific users through filters and views that can be configured individually. Tivoli Netcool/OMNIbus has automation functions that can perform intelligent processing on managed alerts.

4.5.2 IBM System Networking Element Manager solution requirements

IBM System Networking Element Manager V6.1 is a software product that is distributed as a virtual appliance. A virtual software appliance requires a hypervisor to enable it to execute. SNEM V6.1 supports the following hypervisors:

- Linux Kernal-based virtual machine (KVM)
- VMware ESX/ESXi

Both KVM and ESX are hypervisors that support SNEM V6.1 as a virtual software appliance. Installation of the ESX version and KVM are independent and mutually exclusive and you must choose the version based on the type of hypervisor you are using.

More information

For more information about the SNEM, see the following resources:

- *IBM SNEM 6.1 Solution Getting Started Guide*:

  http://www-01.ibm.com/support/docview.wss?uid=isg3T7000471&aid=1

- *IBM SNEM 6.1 User Guide*:

  http://www-01.ibm.com/support/docview.wss?uid=isg3T7000473&aid=1

- *IBM SNEM 6.1 Release Notes Changes*:

  http://www-01.ibm.com/support/docview.wss?uid=isg3T7000474&aid=1

- *IBM System Networking Element Manager Solution Device Support List (6.1)*:

  http://www-01.ibm.com/support/docview.wss?uid=isg3T7000474

- Quick Start Guide for installing and running KVM:

  http://publib.boulder.ibm.com/infocenter/lnxinfo/v3r0m0/topic/liaai/kvminstall/kvminstall_pdf.pdf

5

IBM System Networking RackSwitch implementation

This chapter provides information and instructions for implementing the IBM System Networking 10Gb Top-of-Rack family switches, RackSwitch G8264R/F and RackSwitch G8124. Using the reference architecture described in Chapter 3, “Reference architectures” on page 107, this chapter presents a step by step guide for implementing and configuring the most important functions implemented in IBM Networking OS. It is not meant to cover all the available features in the operating system. Instead, it is strictly for to the proposed test architecture.

The goal of this chapter is that the final result is a fully functional network, able to prove the operation of the implemented features.

A reader with the same (or equivalent) equipment used in the reference architecture at his disposal should be able to replicate the described configuration by following the steps presented in this chapter, and arrive at the same result.

This chapter cover implementation aspects that pertain to OSI Layers 1 - 3, as follows:

- Layer 1: Configuration, information, and statistics commands related to Layer 1 operation. Port configuration in terms of speed and duplex, link status, errors, and so on.
- Layer 2: Configuration, information, and statistics commands related to Layer 2 operation. VLANs, ports and trunking, Spanning Tree Protocol (STP), Quality of Service (QoS), and high availability mechanisms.
- Layer 3: Configuration, information, and statistics commands related to Layer 3 operation. Basic IP routing, dynamic routing protocols, high availability mechanisms (Virtual Router Redundancy Protocol (VRRP)), IPv6, and so on.

© Copyright IBM Corp. 2012. All rights reserved.

5.1 Layer 1 implementation

This section describes Layer 1 related configuration and verification information for the implemented reference architecture.

We describe the following topics:

- Network topology for Layer 1 configuration
- Configuration of the port settings

5.1.1 Network topology for Layer 1 configuration

This section describes the Layer 1 implementation of the reference architecture. Figure 5-1shows the physical connections of the lab equipment that we used to demonstrate the examples in this chapter.

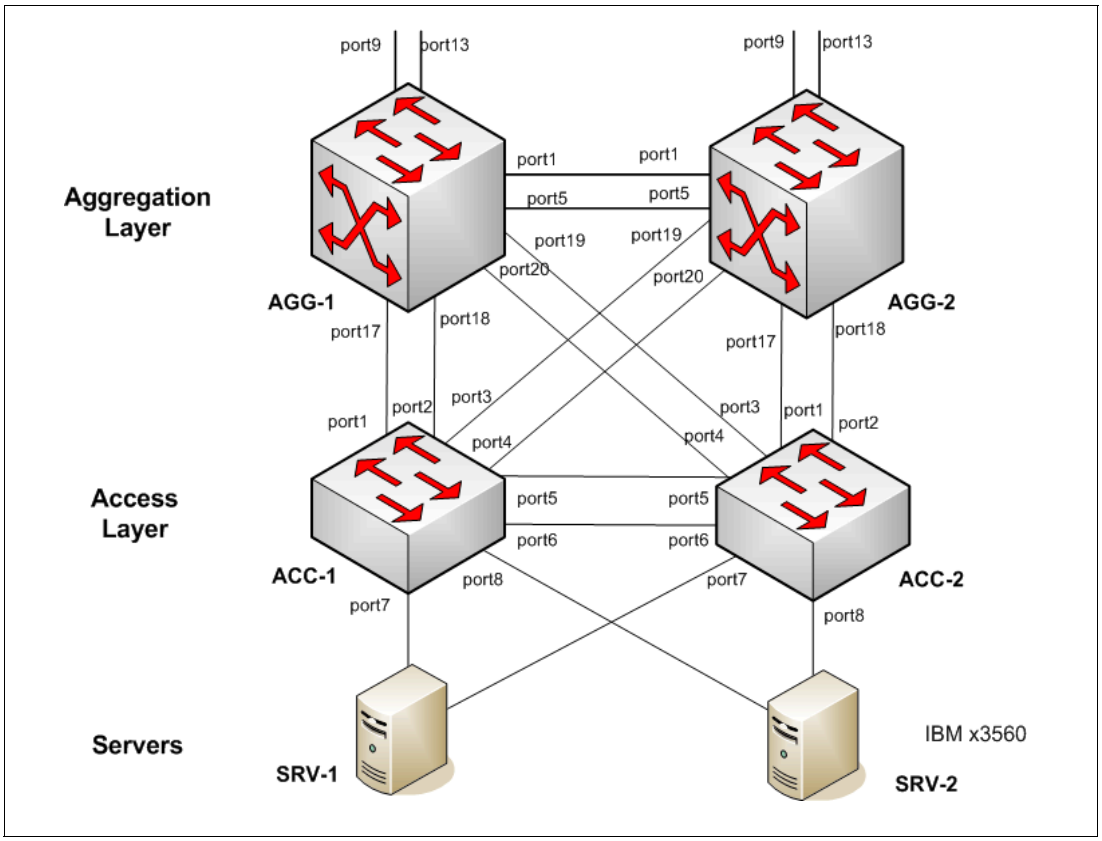

Figure 5-1 Layer 1 topology

All the equipment used in the reference architecture were running IBM Networking Operating System V6.8 (previously know as BLADEOS), and the configuration, statistics, and information commands were tested for each model of the switch. The command syntax and output are related to the OS version, so the syntax and outputs are applicable for all the RackSwitch models used in the lab setup.

Any model-specific difference in terms of software configuration is documented.

5.1.2 Port settings configuration

The physical connections details are provided in the Chapter 3, “Reference architectures” on page 107. Most Layer 1 aspects do not involve any configuration. In this section, we provide some useful commands for ports verification, that is, to make sure all the links are up and running before proceeding to upper layer configuration.

Additional commands and details for Layer 1 configuration can be found in the technical documentation listed in 5.5, “More information” on page 238

Port link configuration

IBM RackSwitch switches include a factory default configuration that enables interfaces with the following link settings:

- In all the copper Gigabit Ethernet interfaces:
  - Auto-negotiation is set.
  - The speed for 10/100/1000 RJ45 (copper) Gigabit Ethernet interfaces is set to auto, so that the interface can operate at 10 Mbps, 100 Mbps, or 1 Gbps. The link operates at the highest possible speed, depending on the capabilities of the remote end. If the speed is manually set to 1 Gbps, the duplex operation is automatically set to full.
  - The duplex mode is set to auto.
  - The flow control is set to none.
- In all the fiber Gigabit Ethernet interfaces:
  - No auto-negotiation is set.
  - The speed is set to 1 Gbps.
  - THe duplex mode is set to full.
  - The flow control is set to none.
- In all the fiber 10 Gigabit Ethernet interfaces:
  - No auto-negotiation is set.
  - The speed is set to 10 Gbps.
  - The duplex mode is set to full.
  - The flow control is set to none.

All the ports used in this implementation are 10 Gbps or 40 Gbps and default to no auto-negotiation / full-duplex. There is no need for speed or duplex configuration.

The port configuration-related commands are listed in this section.

To change the default link parameters on Gigabit Ethernet interfaces of the RackSwitches, enter one of the following interface/portchannel configuration level commands:

- **RackSwitch(config-if)#speed {10|100|1000|auto}**
- **RackSwitch(config-if)#duplex {full|half|any}**
- **RackSwitch(config-if)#[no] flowcontrol {receive|send|both}**
- **RackSwitch(config-if)#[no] auto**

Interface parameters: The default interface parameters are not listed in the configuration.

To temporarily disable a port, run the following command:

```
RackSwitch#interface port <Port alias or number> shutdown
```

Configuring QSFP+

The G8264 RackSwitch is equipped with QFSP+ ports that can operate either in 10 GbE or 40 GbE mode.

Important: Changing the QSFP+ operation mode requires a reboot.

The QSFP+ ports numbering depends on the mode of operation (10 GbE or 40 GbE) (Table 5-1)>

Table 5-1 QSFP+ ports numbering

Physical port number	40 GbE mode	10 GbE mode
Port 1	Port 1	Ports 1 - 4
Port 5	Port 5	Ports 5 - 8
Port 9	Port 9	Ports 9 - 12
Port 13	Port 13	Ports 13 - 16

To change the QSFP+ port mode, complete the following steps:

1. Run **`show boot qsfp-port-modes`** to display the current port mode for the QSFP+ ports (Example 5-1).

Example 5-1 Show the QSFP+ port mode

```
AGG1#show boot qsfp-port-modes
QSFP ports booted configuration:
  Port 1, 2, 3, 4 - 10G Mode
  Port 5, 6, 7, 8 - 10G Mode
  Port 9, 10, 11, 12 - 10G Mode
  Port 13, 14, 15, 16 - 10G Mode

QSFP ports saved configuration:
  Port 1, 2, 3, 4 - 10G Mode
  Port 5, 6, 7, 8 - 10G Mode
  Port 9, 10, 11, 12 - 10G Mode
  Port 13, 14, 15, 16 - 10G Mode
AGG1#
```

2. Change the port mode to 40 GbE by selecting the physical port number by using either an individual port or a list of QSFP+ ports (Example 5-2).

Example 5-2 Change the QSFP+ port mode to 40 GbE

```
AGG1(config)#boot qsfp-40Gports 1,5
```

3. Verify the configuration changes by running the command shown in Example 5-3.

Example 5-3 Verify the configuration change of the QSFP+ port mode to 40 GbE

```
AGG1#show boot qsfp-port-modes
QSFP ports booted configuration:
Port 1, 2, 3, 4 - 10G Mode
Port 5, 6, 7, 8 - 10G Mode
Port 9, 10, 11, 12 - 10G Mode
Port 13, 14, 15, 16 - 10G Mode
```

```
QSFP ports saved configuration:
Port 1 - 40G Mode
Port 5 - 40G Mode
Port 9, 10, 11, 12 - 10G Mode
Port 13, 14, 15, 16 - 10G Mode
AGG1#
```

4. Reset the RackSwitch by running the command shown in Example 5-4.

Example 5-4 Reset the switch

```
AGG1#reload
```

5. Verify the current operation mode of the QSFP+ ports, as shown in Example 5-5.

Example 5-5 Verify the QSFP+ ports operation mode after reboot

```
AGG-1#sh boot qsfp-port-modes
QSFP ports booted configuration:
  Port 1 - 40G Mode
  Port 5 - 40G Mode
  Port 9, 10, 11, 12 - 10G Mode
  Port 13, 14, 15, 16 - 10G Mode

QSFP ports saved configuration:
  Port 1 - 40G Mode
  Port 5 - 40G Mode
  Port 9, 10, 11, 12 - 10G Mode
  Port 13, 14, 15, 16 - 10G Mode
AGG-1#
```

6. Use the **no** form of the command to reset all ports to 10 GbE mode, as shown in Example 5-6.

Example 5-6 Revert the ports to the 10 GbE configuration

```
AGG1(config)#no boot qsfp-40Gports 1,5,9
```

Layer 1 verification

Use the commands described in this section to verify Layer 1 operation of the switch.

Run **show interface transceiver** to display the installed transceivers type, part number, serial number, laser type, and status. Example 5-7 shows the command's output.

Example 5-7 Verification of the installed transceivers

```
AGG-1#show interface transceiver

Name             TX  RXLos TXFlt Volts DegsC TXuW  RXuW  Media   Laser  Approval
---------------- --- ----- ----- ----- ----- ----- ----- ------- ------ ----------
 1 QSFP+ 1       N/A LINK  -N/A-  -.-- --.-  ---.- ---.- 1m QDAC -N/A-  Accepted
         Amphenol         Part:594090001        Date:101120 S/N:APF10460010015
 5 QSFP+ 2       N/A LINK  -N/A-  -.-- --.-  ---.- ---.- 1m QDAC -N/A-  Accepted
         Amphenol         Part:594090001        Date:110125 S/N:APF11030010006
 9 Q10G 3.A      < NO Device Installed >
10 Q10G 3.B      < NO Device Installed >
```

```
11 Q10G 3.C      < NO Device Installed >
12 Q10G 3.D      < NO Device Installed >
13 Q10G 4.A      < NO Device Installed >
14 Q10G 4.B      < NO Device Installed >
15 Q10G 4.C      < NO Device Installed >
16 Q10G 4.D      < NO Device Installed >
17 SFP+  1       N/A LINK  -N/A-  -.-- --.-  ---.- ---.- 1m DAC  -N/A-  Accepted
         BLADE NETWORKS   Part:BN-SP-CBL-1M     Date:091110 S/N:APF09460020129
18 SFP+  2       N/A LINK  -N/A-  -.-- --.-  ---.- ---.- 1m DAC  -N/A-  Accepted
         BLADE NETWORKS   Part:BN-SP-CBL-1M     Date:091106 S/N:APF09450020298
19 SFP+  3       N/A LINK  -N/A-  -.-- --.-  ---.- ---.- 1m DAC  -N/A-  Accepted
         BLADE NETWORKS   Part:BN-SP-CBL-1M     Date:091105 S/N:APF09450020785
20 SFP+  4       N/A LINK  -N/A-  -.-- --.-  ---.- ---.- 1m DAC  -N/A-  Accepted
         BLADE NETWORKS   Part:BN-SP-CBL-1M     Date:091106 S/N:APF09450020645
21 SFP+  5       Ena LINK  no     3.37 32.0  567.9 543.9 SR SFP+ 850nm  Approved
         Blade Network    Part:BN-CKM-SP-SR     Date:100606 S/N:AA1022A4GHJ
22 SFP+  6       Ena LINK  no     3.30 31.5  570.5 616.6 SR SFP+ 850nm  Approved
         Blade Network    Part:BN-CKM-SP-SR     Date:100505 S/N:AA1018A3R8K
23 SFP+  7       < NO Device Installed >
24 SFP+  8       < NO Device Installed >
25 SFP+  9       < NO Device Installed >
26 SFP+ 10       < NO Device Installed >
27 SFP+ 11       < NO Device Installed >
28 SFP+ 12       < NO Device Installed >
29 SFP+ 13       < NO Device Installed >
30 SFP+ 14       < NO Device Installed >
31 SFP+ 15       < NO Device Installed >
32 SFP+ 16       < NO Device Installed >
33 SFP+ 17       < NO Device Installed >
34 SFP+ 18       < NO Device Installed >
35 SFP+ 19       < NO Device Installed >
36 SFP+ 20       < NO Device Installed >
37 SFP+ 21       < NO Device Installed >
38 SFP+ 22       < NO Device Installed >
39 SFP+ 23       < NO Device Installed >
40 SFP+ 24       < NO Device Installed >
41 SFP+ 25       < NO Device Installed >
42 SFP+ 26       < NO Device Installed >
43 SFP+ 27       < NO Device Installed >
44 SFP+ 28       < NO Device Installed >
45 SFP+ 29       < NO Device Installed >
46 SFP+ 30       < NO Device Installed >
47 SFP+ 31       < NO Device Installed >
48 SFP+ 32       < NO Device Installed >
49 SFP+ 33       < NO Device Installed >
50 SFP+ 34       < NO Device Installed >
51 SFP+ 35       < NO Device Installed >
52 SFP+ 36       < NO Device Installed >
53 SFP+ 37       < NO Device Installed >
54 SFP+ 38       < NO Device Installed >
55 SFP+ 39       < NO Device Installed >
56 SFP+ 40       < NO Device Installed >
57 SFP+ 41       < NO Device Installed >
58 SFP+ 42       < NO Device Installed >
59 SFP+ 43       < NO Device Installed >
```

```
60 SFP+ 44       < NO Device Installed >
61 SFP+ 45       < NO Device Installed >
62 SFP+ 46       < NO Device Installed >
63 SFP+ 47       < NO Device Installed >
64 SFP+ 48       < NO Device Installed >
AGG-1#
```

Run **show interface status** or **show interface link** to display information about link, duplex, speed, and flow control. Example 5-8 shows the commands' output.

Example 5-8 Interface status verification

```
AGG-1#show interface status
------------------------------------------------------------------
Alias   Port   Speed    Duplex     Flow Ctrl      Link
------- ----   -----   --------  --TX-----RX--   ------
1        1      40000     full      no     no       up
2        2      10000     full      no     no      down
3        3      10000     full      no     no      down
4        4      10000     full      no     no      down
5        5      40000     full      no     no       up
6        6      10000     full      no     no      down
7        7      10000     full      no     no      down
8        8      10000     full      no     no      down
9        9      10000     full      no     no      down
10       10     10000     full      no     no      down
11       11     10000     full      no     no      down
12       12     10000     full      no     no      down
13       13     10000     full      no     no      down
14       14     10000     full      no     no      down
15       15     10000     full      no     no      down
16       16     10000     full      no     no      down
17       17     10000     full      no     no       up
18       18     10000     full      no     no       up
19       19     10000     full      no     no       up
20       20     10000     full      no     no       up
21       21     10000     full      no     no       up
22       22     10000     full      no     no       up
23       23     10000     full      no     no      down
24       24     1G/10G    full      no     no      down
25       25     1G/10G    full      no     no      down
26       26     1G/10G    full      no     no      down
27       27     1G/10G    full      no     no      down
28       28     1G/10G    full      no     no      down
29       29     1G/10G    full      no     no      down
30       30     1G/10G    full      no     no      down
31       31     1G/10G    full      no     no      down
32       32     1G/10G    full      no     no      down
33       33     1G/10G    full      no     no      down
34       34     1G/10G    full      no     no      down
35       35     1G/10G    full      no     no      down
36       36     1G/10G    full      no     no      down
37       37     1G/10G    full      no     no      down
38       38     1G/10G    full      no     no      down
39       39     1G/10G    full      no     no      down
40       40     1G/10G    full      no     no      down
```

```
41       41     1G/10G    full      no     no      down
42       42     1G/10G    full      no     no      down
43       43     1G/10G    full      no     no      down
44       44     1G/10G    full      no     no      down
45       45     1G/10G    full      no     no      down
46       46     1G/10G    full      no     no      down
47       47     1G/10G    full      no     no      down
48       48     1G/10G    full      no     no      down
49       49     1G/10G    full      no     no      down
50       50     1G/10G    full      no     no      down
51       51     1G/10G    full      no     no      down
52       52     1G/10G    full      no     no      down
53       53     1G/10G    full      no     no      down
54       54     1G/10G    full      no     no      down
55       55     1G/10G    full      no     no      down
56       56     1G/10G    full      no     no      down
57       57     1G/10G    full      no     no      down
58       58     1G/10G    full      no     no      down
59       59     1G/10G    full      no     no      down
60       60     1G/10G    full      no     no      down
61       61     1G/10G    full      no     no      down
62       62     1G/10G    full      no     no      down
63       63     1G/10G    full      no     no      down
64       64     1G/10G    full      no     no      down
MGT      65      1000     full     yes    yes       up
AGG-1#
```

Run **show interface counters** to display traffic statistics for the switch ports. Example 5-9 shows the command's output.

Example 5-9 Display interface traffic statistics

```
AGG-1#show interface counters
------------------------------------------------------------------
Interface statistics for port 1:
                     ifHCIn Counters       ifHCOut Counters
Octets:                     60746287              45114475
UcastPkts:                     32748                  942
BroadcastPkts:                   620                 5020
MulticastPkts:                739219                551871
FlowCtrlPkts:                      0                     0
PriFlowCtrlPkts:                   0                     0
Discards:                          0                     0
Errors:                            0                     0

Ingress Discard reasons for port 1:
VLAN Discards:                     0
Empty Egress Portmap:              0
Filter Discards:                   0
Policy Discards:                   0
Non-Forwarding State:              0
IBP/CBP Discards:                  0

------------------------------------------------------------------
Interface statistics for port 2:
                     ifHCIn Counters       ifHCOut Counters
```

```
Octets:                          0                    0
UcastPkts:                       0                    0
BroadcastPkts:                   0                    0
MulticastPkts:                   0                    0
FlowCtrlPkts:                    0                    0
PriFlowCtrlPkts:                 0                    0
Discards:                        0                    0
Errors:                          0                    0

Ingress Discard reasons for port 2:
VLAN Discards:                   0
Empty Egress Portmap:            0
Filter Discards:                 0
Policy Discards:                 0
Non-Forwarding State:            0
IBP/CBP Discards:                0

----------------------------------------------------------------
Press q to quit, any other key to continue...
```

5.2 Layer 2

This section refers to basic Layer 2 configuration for all Top-of-Rack switches referred to in this publication. Any differences between models regarding the availability of certain software features or command syntax are highlighted.

All the configurations presented in this chapter were implemented by using IBM Networking Operating System V6.8 (previously know as BLADEOS) software installed in the reference architecture switches. Configuration steps and examples are presented for each command on selected equipment. We assume that the template can be easily replicated for the remaining ports and equipment according to the reference architecture. For the architecture planning details, VLAN numbers and assignment, trunks, and so on, see Chapter 3, “Reference architectures” on page 107.

Not all the Layer 2 functions available in IBM Networking OS V6.8 are covered in this section. For an extensive list of features and configuration guidelines, see the documentation links listed in 5.5, “More information” on page 238.

The following topics are described in this section:

- VLANs
- Ports and trunking
- Spanning Tree Protocol
- Quality of Service

5.2.1 VLANs

For information about the VLAN-related configuration that is applied to the reference architecture switches, see Chapter 3, “Reference architectures” on page 107.

The configuration topics described in this section are:

- VLANs and Port VLAN ID Numbers
- VLAN Tagging
- Protocol-based VLANs
- Private VLANs

VLANs and Port VLAN ID numbers

Here we show some basic switching configuration, such as configuring a VLAN, assigning a port to a VLAN, and configuring protocol-based VLANs and private VLANs.

The RackSwitch G8264 and RackSwitch G8124 switches support up to 1024 VLANs per switch. Even though the maximum number of VLANs supported at any time is 1024, each can be identified with any number 1 - 4094.

By default, all data ports are members of VLAN 1, so configure only those ports that belong to other VLANs. VLAN 4095 is used by the management network, which includes the management port.

VLANs definition and port assignment configuration steps

To define the VLANs and configure the port assignment, complete the following steps:

1. Define the VLANs by using the commands shown in Example 5-10.

Example 5-10 VLAN definition

```
AGG-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-1(config)#vlan 100

VLAN number 100 with name "VLAN 100" created.

VLAN 100 was assigned to STG 100.
AGG-1(config-vlan)#name AGG
1-AGG2
AGG-1(config-vlan)#enable
AGG-1(config-vlan)#
AGG-1(config-vlan)#vlan 101

VLAN number 101 with name "VLAN 101" created.

VLAN 101 was assigned to STG 101.
AGG-1(config-vlan)#name AGG1-ACC1
AGG-1(config-vlan)#enable
AGG-1(config-vlan)#
AGG-1(config-vlan)#vlan 103

VLAN number 103 with name "VLAN 103" created.

VLAN 103 was assigned to STG 103.
AGG-1(config-vlan)#name AGG1-ACC2
AGG-1(config-vlan)#enable
AGG-1(config-vlan)#
AGG-1(config-vlan)#end
AGG-1#
```

Important: Example 5-10 on page 164 is a template. Complete the configuration of the VLANs for all the reference architecture switches according to the details provided in the Chapter 3, “Reference architectures” on page 107.

To show the VLAN configuration on the switch, run **show vlan [information]** (Example 5-11).

Example 5-11 show vlan command output

```
AGG-1#show vlan
VLAN                Name                Status            Ports
----  --------------------------------  ------  ------------------------
1     Default VLAN                      ena     1-64
100   AGG1-AGG2                         ena     empty
101   AGG1-ACC1                         ena     empty
103   AGG1-ACC2                         ena     empty
4095  Mgmt VLAN                         ena     MGT
AGG-1#
```

The Ports column is empty, as no member port is allocated.

Use the **information** option for more detailed output.

2. Assign ports to a VLAN. If you want to assign a port to a VLAN, run **member** at the VLAN configuration level.

   Before you assign ports to VLANs, define and name the VLANs. However, if you assign a port to a non-existent VLAN, the OS automatically create one, according to the PVID defined for the port, and allocates a Spanning Tree Group for it. The VLAN port assignment configuration is shown in Example 5-12.

Example 5-12 VLAN port assignment configuration

```
AGG-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-1(config)#vlan 100
AGG-1(config-vlan)#member 1,5
Port 1 is an UNTAGGED port and its PVID is changed from 1 to 100
Port 5 is an UNTAGGED port and its PVID is changed from 1 to 100
AGG-1(config-vlan)#
AGG-1(config-vlan)#
AGG-1(config-vlan)#vlan 101
AGG-1(config-vlan)#member 17,18
Port 17 is an UNTAGGED port and its PVID is changed from 1 to 101
Port 18 is an UNTAGGED port and its PVID is changed from 1 to 101
AGG-1(config-vlan)#
AGG-1(config-vlan)#
AGG-1(config-vlan)#vlan 103
AGG-1(config-vlan)#member 19-20
Port 19 is an UNTAGGED port and its PVID is changed from 1 to 103
Port 20 is an UNTAGGED port and its PVID is changed from 1 to 103
AGG-1(config-vlan)#
AGG-1(config-vlan)#
```

The resulting PVID assignment configuration is shown in Example 5-13.

Example 5-13 PVID assignment configuration

```
interface port 1
        name "AGG1-AGG2"
        pvid 100
        exit
!
interface port 5
        name "AGG1-AGG2"
        pvid 100
        exit
!
interface port 17
        name "AGG1-ACC1"
        pvid 101
        exit
!
interface port 18
        name "AGG1-ACC1"
        pvid 101
        exit
!
interface port 19
        name "AGG1-ACC2"
        pvid 103
        exit
!
interface port 20
        name "AGG1-ACC2"
        pvid 103
        exit
!
vlan 1
        member 2-4,6-16,21-64
        no member 1,5,17-20
!
vlan 100
        enable
        name "AGG1-AGG2"
        member 1,5
!
vlan 101
        enable
        name "AGG1-ACC1"
        member 17-18
!
vlan 103
        enable
        name "AGG1-ACC2"
        member 19-20
```

Important: Example 5-13 on page 166 is a template. Complete the configuration of the ports and VLAN assignment for all the reference architecture switches according to the details provided in Chapter 3, “Reference architectures” on page 107.

Run **show vlan** to show the VLANs member ports (Example 5-14).

Example 5-14 Port allocation shown by using the show vlan command

```
AGG-1#show vlan
VLAN                Name                Status            Ports
----  --------------------------------  ------  ------------------------
1     Default VLAN                      ena     2-4 6-16 21-64
100   AGG1-AGG2                         ena     1 5
101   AGG1-ACC1                         ena     17 18
103   AGG1-ACC2                         ena     19 20
4095  Mgmt VLAN                         ena     MGT
AGG-1#
```

VLAN tagging

IBM Networking OS software supports 802.1Q VLAN tagging, which provides standards-based VLAN support for Ethernet systems. The default configuration settings for RackSwitch switches have all the ports set as untagged members of VLAN 1 with all ports configured as PVID = 1.

For a detailed description about tagging and terminology, see Chapter 3, “Reference architectures” on page 107.

VLAN tagging is required only on ports that are connected to other switches or on ports that connect to tag-capable endpoints, such as servers with VLAN-tagging adapters.

Tagging is enabled only on the trunk that connects ACC-1 and ACC-2 access switches in the reference architecture. For more details, see Figure 3-2 on page 111 and Table 3-3 on page 111.

To enable tagging for a port, run **tagging** in interface configuration mode (Example 5-15).

Example 5-15 Tagging configuration

```
ACC-2#configure terminal
ACC-2(config-if)#interface port 5,6
ACC-2(config-if)#tagging
ACC-2(config-if)#^Z
ACC-2#
The port configuration will look like this:
interface port 5
        name "ACC1-ACC2"
        tagging
        exit
!
interface port 6
        name "ACC1-ACC2"
        tagging
        exit
```

To allow communication over a tagging enabled connection, the end ports of the switch must be declared members of the required VLANs to be transported over the link.

Ports 5 and 6 on ACC-1 and ACC-2 are configured members of VLAN 10 and VLAN 20:

```
ACC-2#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-2(config)#vlan 10,20
ACC-2(config-vlan)#member 5,6
ACC-2(config-vlan)#^Z
ACC-2#
```

Important: This example is a template. Complete the configuration of port tagging for all the reference architecture switches according to the details provided in Chapter 3, “Reference architectures” on page 107.

Information about tagging, port names, allocated VLANs, PVIDs, and other flags are summarized in the **show interface information** command output shown in Example 5-16. Notice the <**y**> flag on the Tag column for ports 5 and 6.

Example 5-16 Interface parameters summary

```
ACC-1#show interface information
Alias   Port Tag RMON Lrn Fld PVID       NAME                 VLAN(s)
------- ---- --- ---- --- --- ----- -------------- --------------------------------
1       1     n   d    e   e   101  AGG1-ACC1      101
2       2     n   d    e   e   101  AGG1-ACC1      101
3       3     n   d    e   e   102  AGG2-ACC1      102
4       4     n   d    e   e   102  AGG2-ACC1      102
5       5     y   d    e   e    10  ACC1-ACC2      10 20
6       6     y   d    e   e    10  ACC1-ACC2      10 20
7       7     n   d    e   e    10  SRV1           10
8       8     n   d    e   e    20  SRV2           20
9       9     n   d    e   e     1                 1
10      10    n   d    e   e     1                 1
11      11    n   d    e   e     1                 1
12      12    n   d    e   e     1                 1
13      13    n   d    e   e     1                 1
14      14    n   d    e   e     1                 1
15      15    n   d    e   e     1                 1
16      16    n   d    e   e     1                 1
17      17    n   d    e   e     1                 1
18      18    n   d    e   e     1                 1
19      19    n   d    e   e     1                 1
20      20    n   d    e   e     1                 1
21      21    n   d    e   e     1                 1
22      22    n   d    e   e     1                 1
23      23    n   d    e   e     1                 1
24      24    n   d    e   e     1                 1
MGTA    25    n   d    e   e  4095                 4095
MGTB    26    n   d    e   e  4095                 4095

* = PVID is tagged.
ACC-1#
```

Protocol-based VLANs

This feature is not part of our reference architecture implementation. However, for completeness, we provide a summary of commands used for configuration and verification in this section.

For more information about the protocols-based VLANs concept, see Chapter 2, “IBM System Networking Switch 10Gb Ethernet switch features” on page 51.

Important: This feature is supported only on RackSwitch G8264.

Consider the following guidelines when you configure protocol-based VLANs:

- Each port can support up to 16 VLAN protocols.
- The RackSwitch G8264 switch can support up to 16 protocols simultaneously.
- Each PVLAN must have at least one port assigned before it can be activated.
- The same port within a port-based VLAN can belong to multiple PVLANs.
- An untagged port can be a member of multiple PVLANs.
- A port cannot be a member of different VLANs with the same protocol association.

Run the following command to configure protocol-based VLAN for a selected VLAN. Run the commands at the VLAN level to configure the frame type and the Ethernet type for the selected protocol.

```
protocol-vlan <1-8> frame-type {ether2|llc|snap} <Ethernet type>
```

The Ethernet type is a 4-digit (16 bit) hex code, such as 0080 (IPv4).

Run **protocol-vlan <1-8> protocol <protocol type>** at the VLAN level to select a predefined protocol. The predefined protocols include:

- decEther2: DEC Local Area Transport
- ipv4Ether2: Internet IP (IPv4)
- ipv6Ether2: IPv6
- ipx802.2: Novell IPX 802.2
- ipx802.3: Novell IPX 802.3
- ipxEther2: Novell IPX
- ipxSnap: Novell IPX SNAP
- netbios: NetBIOS 802.2
- rarpEther2: Reverse ARP
- sna802.2: SNA 802.2
- snaEther2: IBM SNA Service on Ethernet
- vinesEther2: Banyan VINES
- xnsEther2: XNS Compatibility

Run **protocol-vlan <1-8> priority <0-7>** at the VLAN level to configure the priority value for this PVLAN.

Run **[no] protocol-vlan <1-8> member <port alias or number>** at the VLAN level to add or remove a port to the selected PVLAN.

Run **[no] protocol-vlan <1-8> tag-pvlan <port alias or number>** at the VLAN level to define a port that is tagged by the selected protocol on this VLAN.

Run **[no] protocol-vlan <1-8> enable** at the VLAN level to enable/disable the selected protocol on the VLAN.

Run **no protocol-vlan <1-8>** at the VLAN level to delete the selected protocol configuration from the VLAN.

Run **show protocol-vlan <1-8>** at the VLAN level to display current parameters for the selected PVLAN.

Private VLANs

This feature is not part of our reference architecture implementation. However, for completeness, we give a summary of the commands used for configuration and verification in this section.

For more information about the private VLANs concept, see Chapter 2, “IBM System Networking Switch 10Gb Ethernet switch features” on page 51.

The following guidelines apply when configuring Private VLANs:

- The default VLAN 1 cannot be a Private VLAN.
- The management VLAN 4095 cannot be a Private VLAN. Management ports cannot be members of Private VLANs.
- IGMP Snooping must be disabled on isolated VLANs.
- Each secondary port’s (isolated port and community ports) PVID must match its corresponding secondary VLAN ID.
- Ports within a secondary VLAN cannot be members of other VLANs.
- All VLANs that are part of the Private VLAN must belong to the same Spanning Tree Group.

Run the following commands to configure Private VLANs:

- Run **private-vlan type primary** at the VLAN level to configure the VLAN type as a Primary VLAN. A Private VLAN must have only one primary VLAN. The primary VLAN carries unidirectional traffic to ports on the isolated VLAN or to community VLAN.
- Run **private-vlan type community** at the VLAN level to configure the VLAN type as a community VLAN. Community VLANs carry upstream traffic from host ports. A Private VLAN may have multiple community VLANs.
- Run **private-vlan type isolated** at the VLAN level to configure the VLAN type as an isolated VLAN. The isolated VLAN carries unidirectional traffic from host ports. A Private VLAN may have only one isolated VLAN.
- Run **no private-vlan type** to clear the private-VLAN type.
- Run **[no] private-vlan map [<2-4094>]** to configure Private VLAN mapping between a secondary VLAN and a primary VLAN. Enter the primary VLAN ID. Secondary VLANs have the type defined as isolated or community. Use the **no** parameter to remove the mapping between the secondary VLAN and the primary VLAN.
- Run **[no] private-vlan enable** at the VLAN level to enable/disable the private VLAN.
- Run **show private-vlan [<2-4094>]** to display the current parameters for the selected Private VLAN(s).

5.2.2 Ports and trunking

When using port trunk groups between two switches, you can create a virtual link between the switches, which operates with combined throughput levels that depend on how many physical ports are included.

Two trunk types are available:

- Static trunk groups (portchannel)
- Dynamic LACP trunk groups

Regarding the two RackSwitch types:

RackSwitch G8264 A RackSwitch G8264 switch supports up to 64 trunk groups (static and LACP). Each type can contain up to 16 member ports.

RackSwitch G8124 A RackSwitch G8124 switch supports up to 24 trunk groups on the switch (static and LACP). Each type can contain up to eight member ports. Of the available configuration slots, any of them may be used for LACP trunks, although only up to 12 may be used for static trunks. In addition, although up to a total of 24 trunks may be configured and enabled, only a maximum of 16 may be operational at any time.

For example, if you configure and enable 12 static trunks (the maximum), up to four LACP trunks may also be configured and enabled, for a total of 16 operational trunks. If more than 16 trunks are enabled at any time, after the switch establishes the 16th trunk group, any additional trunks are automatically placed in a non-operational state. In this scenario, there is no administrative means to ensure which 16 trunks are selected for operation.

The trunking feature operates according to specific configuration rules. When creating trunks, consider the following rules that determine how a trunk group reacts in any network topology:

- All trunks must originate from one logical device, and lead to one logical destination device. Usually, a trunk connects two physical devices together with multiple links. However, in some networks, a single logical device may include multiple physical devices, such as when switches are configured in a stack, or when using VLAGs. In such cases, links in a trunk can connect to multiple physical devices because they act as one logical device.
- Any physical switch port can belong to only one trunk group.
- Trunking from third-party devices must comply with Cisco EtherChannel technology.
- All ports in a trunk must have the same link configuration (speed, duplex, and flow control), the same VLAN properties, and the same Spanning Tree Protocol, storm control, and ACL configuration. The ports in a trunk should be members of the same VLAN.
- Each trunk inherits its port configuration (speed, flow control, and tagging) from the first member port. As additional ports are added to the trunk, their settings must be changed to match the trunk configuration.
- When a port leaves a trunk, its configuration parameters are retained.
- You cannot configure a trunk member as a monitor port in a port-mirroring configuration.
- Trunks cannot be monitored by a monitor port; however, trunk members can be monitored.

Static trunks

By default, each trunk group is empty and disabled. To define a static trunk group, complete the configuration steps in this section.

Static trunks: Static trunks are configured in the reference architecture on the links between AGG-1 and AGG-2, AGG-1 and ACC-1, AGG-1 and ACC-2, AGG-2 and ACC-1, and AGG-2 and ACC-2.

1. Add physical ports to a trunk group.

   Run **[no] portchannel <number> port <port alias or number>** from the global configuration mode to add or remove ports to or from a trunk group (Example 5-17).

   Example 5-17 Static trunk configuration

```
AGG-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-1(config)#portchannel 1 port 17-18
AGG-1(config)#portchannel 2 port 19,20
AGG-1(config)#portchannel 3 port 1,5
AGG-1(config)#^Z
AGG-1#
```

2. Enable the trunk group.

   Run **[no] portchannel <number> enable** from the global configuration mode to enable or disable trunk groups (Example 5-18).

   Example 5-18 Enable static trunks

```
AGG-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-1(config)#portchannel 1 enable
AGG-1(config)#portchannel 2 enable
AGG-1(config)#portchannel 3 enable
AGG-1(config)#^Z
AGG-1#
```

   You can remove the current trunk group configuration by running **no portchannel <number>**.

3. Configure hashing.

   Traffic in a trunk group is statistically distributed among member ports by using a hash process, where various address and attribute bits from each transmitted frame are recombined to specify the particular trunk port the frame uses.

   The switch can be configured to use various hashing options. To achieve the most even traffic distribution, select options that exhibit a wide range of values for your particular network. Avoid hashing of information that is not present in the expected traffic, or which does not vary.

   Trunk hash parameters are set globally. You can enable one or two parameters, to configure any of the following valid combinations:

   – SMAC (source MAC only)
   – DMAC (destination MAC only)
   – SIP (source IP only)
   – DIP (destination IP only)
   – SIP + DIP (source IP and destination IP)
   – SMAC + DMAC (source MAC and destination MAC)

Syntax: There are command syntax differences between the RackSwitch G8264 and RackSwitch G8124 switches. Both sets of commands are presented in Table 5-2 on page 173.

Table 5-2 Trunk hashing configuration commands

Hash options	RackSwitch G8264 command	RackSwitch G8124 command
Layer 2		
SMAC	`portchannel thash l2thash l2-source-mac-address`	`portchannel hash source-mac-address`
DMAC	`portchannel thash l2thash l2-destination-mac-address`	`portchannel hash destination-mac-address`
SMAC+DMAC	`portchannel thash l2thash l2-source-destination-mac`	`portchannel hash source-destination-mac`
Layer 3		
SIP	`portchannel thash l3thash l3-source-ip-address`	`portchannel hash source-ip-address`
DIP	`portchannel thash l3thash l3-destination-ip-address`	`portchannel hash destination-ip-address`
SIP+DIP	`portchannel thash l3thash l3-source-destination-ip`	`portchannel hash source-destination-ip`
L3 use L2	`portchannel thash l3thash l3-use-l2-hash`	
Other options		
Ingress port	`portchannel thash ingress`	
L4 port	`portchannel thash L4port`	

Hashing options: The RackSwitch G8264 **`ingress`** and **`L4 port`** hashing options are disabled by default.

In Example 5-19, source MAC address hashing was enabled for all switches and for RackSwitch G8264, and ingress port hashing was also configured.

Example 5-19 Configure trunk hashing

```
AGG-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-1(config)#portchannel thash l2thash l2-source-mac-address
AGG-1(config)#portchannel thash ingress
AGG-1(config)#^Z
AGG-1#

ACC-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-1(config)#portchannel hash source-mac-address
ACC-1(config)#^Z
ACC-1#
```

Run **show portchannel hash** to verify the global hash parameters shown in Example 5-20.

Example 5-20 Display trunk hashing parameters

```
AGG-1#show portchannel hash
Current L2 trunk hash settings:
    smac
Current L3 trunk hash settings:
    sip dip
Current ingress port hash: enabled
Current L4 port hash: disabled
AGG-1#


ACC-1#show portchannel hash
Current  Trunk Hash settings:
    smac
ACC-1#
```

IP has: Source IP and destination IP hash is enabled by default on RackSwitch G8264 switches.

4. Verify the trunk group configuration.

   Run the commands in this section to verify the trunk configuration and status.

   Verify the trunk group status by running **show portchannel [information]** or **show portchannel <number> [information]**. The output of this command is shown in Example 5-21.

Example 5-21 Display trunk group status information

```
AGG-1#show portchannel information
PortChannel 1: Enabled
Protocol - Static
Port State:
     17: STG 101 forwarding
     18: STG 101 forwarding

PortChannel 2: Enabled
Protocol - Static
Port State:
     19: STG 103 forwarding
     20: STG 103 forwarding

PortChannel 3: Enabled
Protocol - Static
Port State:
      1: STG 100 forwarding
      5: STG 100 forwarding

AGG-1#
```

Important: Example 5-21 is a template. Complete the configuration of trunks for all the reference architecture switches according to the details provided in Chapter 3, “Reference architectures” on page 107.

Verify the trunk group parameters by running **show portchannel <number>** **(**Example 5-22).

Example 5-22 Display trunk group parameters

```
AGG-1#show portchannel 1
Protocol - Static
Current settings: enabled
    ports:  17, 18
Current L2 trunk hash settings:
    smac
Current L3 trunk hash settings:
    sip dip
Current ingress port hash: enabled
Current L4 port hash: disabled

AGG-1#
```

Dynamic trunks

Link Aggregation Control Protocol (LACP) is an IEEE 802.3ad standard for grouping several physical ports into one logical port (known as a dynamic trunk group or Link Aggregation group) with any device that supports the standard. See IEEE 802.3ad-2002 for a full description of the standard.

The 802.3ad standard allows standard Ethernet links to form a single Layer 2 link by using the LACP. Link aggregation is a method of grouping physical link segments of the same media type and speed in full duplex, and treating them as if they were part of a single, logical link segment. If a link in a LACP trunk group fails, traffic is reassigned dynamically to the remaining links of the dynamic trunk group.

A port's Link Aggregation Identifier (LAG ID) determines how the port can be aggregated. The LAG ID is constructed mainly from the system ID and the port's admin key, as follows:

System ID An integer value based on the switch's MAC address and the system priority assigned in the CLI.

Admin key A port's Admin key is an integer value (1 - 65535) that you can configure in the CLI. Each switch port that participates in the same LACP trunk group must have the same *admin key* value. The Admin key is *local significant*, which means the partner switch does not need to use the same Admin key value.

LACP automatically determines which member links can be aggregated and then aggregates them. It provides for the controlled addition and removal of physical links for the link aggregation.

Each port on the switch can have one of the following LACP modes:

Off (default) The user can configure this port in a regular static trunk group.

Active The port can form an LACP trunk. This port sends LACPDU packets to partner system ports.

Passive The port can form an LACP trunk. This port responds only to the LACPDU packets sent from an LACP active port.

Important:

- When the system is initialized, all ports by default are in LACP off mode and are assigned unique admin keys.
- A dynamic trunk is configured in the reference architecture on the link between ACC-1 and ACC-2

Follow these steps to configure and activate LACP trunks:

1. Configure the global LACP parameters by running the following commands:
   a. Run **lacp system-priority <1-65535>** to define the priority value for the switch. Lower numbers provide higher priority. The default value is 32768. We use the default value in our example.
   b. Run **lacp timeout {short|long}** to define the timeout period before invalidating LACP data from a remote partner. Choose short (3 seconds) or long (90 seconds). The default value is long, which is used in our example.

   Preferred practice: Use a timeout value of long to reduce LACPDU processing. If your RackSwitch G8124 switche's processor utilization rate remains at 100% for periods of 90 seconds or more, consider using static trunks instead of LACP.

   a. Run **no lacp <1-65535>** to delete the selected LACP trunk, based on its admin key. This command is equivalent to disabling LACP on each of the ports configured with the same admin key.
2. Configure the LACP ports by running the following commands:
   a. Set the LACP mode for the ports by running **lacp mode {off|active|passive}**.

      You must set the port's LACP mode to active to activate LACP negotiation. You can set other port's LACP mode to passive, to reduce the amount of LACPDU traffic at the initial trunk-forming stage.

      The LACP ports mode configuration is shown in Example 5-23.

      Example 5-23 LACP ports mode configuration

```
ACC-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-1(config)#interface port 5,6
ACC-1(config-if)#lacp mode active
ACC-1(config-if)#^Z
ACC-1#
```

   b. Set the LACP priority value for a selected port by running **lacp priority <1-65535>**. Lower numbers provide higher priority. The default value is 32768. The LACP port priority configuration is shown in Example 5-24.

      Example 5-24 LACP port priority configuration

```
ACC-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-1(config)#interface port 5
ACC-1(config-if)#lacp priority 16384
ACC-1(config-if)#^Z
ACC-1#
```

c. Set the admin key for the selected ports by running **lacp key <1-65535>**. Only ports with the same adminkey and operkey (the operational state is generated internally) can form a LACP trunk group. The LACP admin key configuration is shown in Example 5-25.

Example 5-25 LACP admin key configuration

```
ACC-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-1(config)#interface port 5,6
ACC-1(config-if)#lacp key 3
ACC-1(config-if)#^Z
ACC-1#
```

3. Verify the LACP configuration and operation by running the following commands:

a. Run **show lacp** to show the current LACP configuration (Example 5-26).

Example 5-26 Global LACP configuration

```
ACC-1#show lacp
Current LACP system ID: 08:17:f4:34:4d:00
Current LACP system Priority: 32768
Current LACP timeout scale: long

Current LACP params for 5:  active, Priority 16384, Admin Key 3, Min-Links 1

Current LACP params for 6:  active, Priority 32768, Admin Key 3, Min-Links 1

ACC-1#
```

b. Run **show lacp information** to show the current LACP parameters for all ports (Example 5-27).

Example 5-27 LACP information for all ports

```
ACC-1#show lacp information
port    mode    adminkey  operkey  selected   prio  aggr  trunk  status  minlinks
--------------------------------------------------------------------------------
1       off            1        1     no     32768    --     --    --        1
2       off            2        2     no     32768    --     --    --        1
3       off            3        3     no     32768    --     --    --        1
4       off            4        4     no     32768    --     --    --        1
5       active         3        3    yes     32768     3     15    up        1
6       active         3        3    yes     32768     3     15    up        1
7       off            7        7     no     32768    --     --    --        1
8       off            8        8     no     32768    --     --    --        1
9       off            9        9     no     32768    --     --    --        1
10      off           10       10     no     32768    --     --    --        1
11      off           11       11     no     32768    --     --    --        1
12      off           12       12     no     32768    --     --    --        1
13      off           13       13     no     32768    --     --    --        1
14      off           14       14     no     32768    --     --    --        1
15      off           15       15     no     32768    --     --    --        1
16      off           16       16     no     32768    --     --    --        1
17      off           17       17     no     32768    --     --    --        1
18      off           18       18     no     32768    --     --    --        1
19      off           19       19     no     32768    --     --    --        1
20      off           20       20     no     32768    --     --    --        1
21      off           21       21     no     32768    --     --    --        1
```

```
22      off          22       22     no     32768    --     --    --        1
23      off          23       23     no     32768    --     --    --        1
24      off          24       24     no     32768    --     --    --        1
ACC-1#
```

c. Run **show lacp aggregator <1-24>** to show aggregation information for the selected admin key (Example 5-28).

Example 5-28 Aggregator information

```
ACC-1#show lacp aggregator 3
Aggregator Id  3
---------------------------------------------
Aggregator MAC address  - 08:17:f4:34:4d:25
Actor System Priority   - 32768
Actor System ID         - 08:17:f4:34:4d:00
Individual              - FALSE
Actor Oper Key          - 3
Partner System Priority - 32768
Partner System ID       - 08:17:f4:34:4c:00
Partner Oper Key        - 3
ready                   - TRUE
Min-Links               - 1
Number of Ports in aggr - 2
index  0    port 5
index  1    port 6
ACC-1#
```

d. Run **show portchannel information** to show the current operating status of both static and dynamic trunks (Example 5-29).

Example 5-29 Show the trunk status

```
ACC-1#sh portchannel information
PortChannel 1: Enabled
Protocol - Static
Port State:
      1: STG 101 forwarding
      2: STG 101 forwarding
PortChannel 2: Enabled
Protocol - Static
Port State:
      3: STG 102 forwarding
      4: STG 102 forwarding
PortChannel 15: Enabled
Protocol - LACP
Port State:
      5: STG 10 forwarding
         STG 20 forwarding
      6: STG 10 forwarding
         STG 20 forwarding
ACC-1#
```

5.2.3 Spanning Tree Protocol

The Spanning Tree Protocol used for the reference architecture is Per-VLAN Rapid Spanning Tree (PVRST). PVRST mode is based on RSTP, which provides rapid Spanning Tree convergence, and allows for multiple Spanning Tree Groups (STGs), with STGs on a per-VLAN basis. PVRST mode is compatible with Cisco R-PVST/R-PVST+ mode.

To simplify switch configuration, VLAN Automatic STG Assignment (VASA) can be used in SPT/PVST+ or PVRST modes. When VASA is enabled, it is no longer necessary to manually assign an STG for each new VLAN. Instead, each newly configured VLAN is automatically assigned its own STG. If an empty STG is not available, the VLAN is automatically assigned to the default VLAN. When a VLAN is deleted, if there is no other VLAN associated with the assigned STG, and the STG is returned to the available pool.

Up to 128 STGs can be configured on the switch. STG 128 is reserved for management.

VASA is disabled by default, but can be enabled or disabled by running **`spanning-tree stg-auto`**:

```
AGG-1(config)#spanning-tree stg-auto
Warning: all VLANs will be assinged to a STG automatically.
AGG-1(config)#
```

VASA applies only to STP/PVST+ and PVRST modes and is ignored in RSTP and MSTP modes. When VASA is enabled, manual STG assignment is still available. The administrator may assign a specific STG to a VLAN by running regular commands, such as the following one:

```
# spanning-tree stp <STG> vlan <vlan ID>
```

When changing to STP/PVST+ or PVSRT mode (either from RSTP or MSTP modes, or when STP is disabled), all existing VLANs are assigned to a unique STG.

For simplicity and consistency of numbering conventions in our example, the VLAN IDs are assigned numbers lower than 127 (as shown in Example 5-10 on page 164) so the VASA can also assign matching STG numbers for each VLAN. The STG configuration is shown in Example 5-30.

Example 5-30 VLANs and STGs configuration

```
vlan 100
        enable
        name "AGG1-AGG2"
        member 1,5
vlan 101
        enable
        name "AGG1-ACC1"
        member 17-18
vlan 103
        enable
        name "AGG1-ACC2"
        member 19-20
spanning-tree stp 100 vlan 100
spanning-tree stp 101 vlan 101
spanning-tree stp 103 vlan 103
```

According to the reference architecture described in Chapter 3, “Reference architectures” on page 107, the links between AGG1 and AGG2, AGG1 and ACC1, AGG1 and ACC2, AGG2 and ACC1, and AGG2 and ACC2 are point-to-point Layer 3 links. The corresponding VLANs do not span across the network, so no loops are formed. The STP runs on the default configuration.

A configuration other than the default configuration is applied only for VLAN 10 and VLAN 20, which contains the hosts in our topology.

The PVRST and VRRP configuration are consistent, so one access switch (ACC-1) forwards traffic for VLAN 10 and the other access switch (ACC-2) forwards traffic for VLAN 20. This configuration provides load balancing between access switches and a more efficient use of network resources by distributing traffic and processing among different equipment and links. The per-VLAN roles of the access switches are shown in Table 5-3.

Table 5-3 Per-VLAN roles of the access switches

VLAN	PVRST primary root	PVRST secondary root	VRRP master	VRRP backup
10	ACC-1 priority 0	ACC-2 priority 4096	ACC-1 priority 105	ACC-2 priority 100
20	ACC-2 priority 0	ACC-1 priority 4096	ACC-2 priority 105	ACC-1 priority 100

Run **spanning-tree stp <STG number> bridge priority <0-65535>** to configure the bridge priority for a specified VLAN (Example 5-31).

Example 5-31 Bridge priority configuration example

```
ACC-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-1(config)#spanning-tree stp 10 bridge priority 0
ACC-1(config)#spanning-tree stp 20 bridge priority 4096
ACC-1(config)#^Z
ACC-1#

Aug  2  9:57:30 ACC-1 ALERT   stg: STG 10, new root bridge


ACC-2#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-2(config)#spanning-tree stp 20 bridge priority 0
ACC-2(config)#spanning-tree stp 10 bridge priority 4096
ACC-2(config)#^Z
ACC-2#
```

The commands appear under the Spanning Tree section of the configuration:

```
spanning-tree stp 10 bridge priority 0
spanning-tree stp 10 vlan 10

spanning-tree stp 20 bridge priority 4096
spanning-tree stp 20 vlan 20
```

For more detailed configuration options for STP (global commands, timers, and ports parameters), see 5.5, “More information” on page 238.

To verify the STP configuration and operation, run the following command with the output from the reference architecture.

The **spanning-tree stp <STG number> [bridge|information]**command shows detailed information about the STP operation. If you do not provide an STG number, the command output displays the STP information for all groups, as shown in Example 5-32. If you want to narrow the output to a specific STG, add the optional parameters.

Example 5-32 show spanning-tree command output

```
ACC-1#show spanning-tree
------------------------------------------------------------------
Pvst+ compatibility mode enabled

------------------------------------------------------------------
Spanning Tree Group 1: On (PVRST)
VLANs:  1

Current Root:            Path-Cost  Port Hello MaxAge FwdDel
 1000 02:1c:73:08:ea:0b     1990      22    2     20     15

Parameters:  Priority  Hello  MaxAge  FwdDel  Aging   Topology Change Counts
              32769      2      20      15     300                 43

   Port       Prio   Cost     State  Role Designated Bridge      Des Port Type
------------- ---- ---------- -----  ---- ---------------------- -------- ----
13             128     2000!   FWD   DESG 8001-08:17:f4:34:4d:00     800d  P2P
22     (pc13)  128     1990!+  FWD   ROOT 1000-02:1c:73:08:ea:0b     8068  P2P
23     (pc14)  128     1990!+  DISC  ALTN 1000-02:1c:73:08:ea:0b     8068  P2P
24               0        0    FWD *
* = STP turned off for this port.
! = Automatic path cost.
+ = Portchannel cost, not the individual port cost.

------------------------------------------------------------------
Spanning Tree Group 10: On (PVRST)
VLANs:  10

Current Root:            Path-Cost  Port Hello MaxAge FwdDel
 000a 08:17:f4:34:4d:00        0       0    2     20     15

Parameters:  Priority  Hello  MaxAge  FwdDel  Aging   Topology Change Counts
                10      2      20      15     300                 10

   Port       Prio   Cost     State  Role Designated Bridge      Des Port Type
------------- ---- ---------- -----  ---- ---------------------- -------- ----
5      (pc15)  128     1990!+  FWD   DESG 000a-08:17:f4:34:4d:00     8029  P2P
! = Automatic path cost.
+ = Portchannel cost, not the individual port cost.

------------------------------------------------------------------
Spanning Tree Group 20: On (PVRST)
VLANs:  20

Current Root:            Path-Cost  Port Hello MaxAge FwdDel
 0014 08:17:f4:34:4c:00     1990       5    2     20     15
```

```
Parameters:  Priority  Hello  MaxAge  FwdDel  Aging   Topology Change Counts
               4116      2      20      15     300                  10

   Port       Prio   Cost     State  Role Designated Bridge      Des Port Type
------------ ---- ---------- -----  ---- ---------------------- -------- ----
5      (pc15)  128     1990!+  FWD   ROOT 0014-08:17:f4:34:4c:00     8027  P2P
! = Automatic path cost.
+ = Portchannel cost, not the individual port cost.

----------------------------------------------------------------------
Spanning Tree Group 101: On (PVRST)
VLANs:  101

Current Root:            Path-Cost  Port Hello MaxAge FwdDel
 8065 08:17:f4:32:c4:00      990       1    2     20     15

Parameters:  Priority  Hello  MaxAge  FwdDel  Aging   Topology Change Counts
               32869      2      20      15     300                   8

   Port       Prio   Cost     State  Role Designated Bridge      Des Port Type
------------ ---- ---------- -----  ---- ---------------------- -------- ----
1      (pc1)   128      990!+  FWD   ROOT 8065-08:17:f4:32:c4:00     8042  P2P
2      (pc1)   128      990!+  FWD   ROOT 8065-08:17:f4:32:c4:00     8042  P2P
! = Automatic path cost.
+ = Portchannel cost, not the individual port cost.

----------------------------------------------------------------------
Spanning Tree Group 102: On (PVRST)
VLANs:  102

Current Root:            Path-Cost  Port Hello MaxAge FwdDel
 8066 08:17:f4:34:4d:00        0       0    2     20     15

Parameters:  Priority  Hello  MaxAge  FwdDel  Aging   Topology Change Counts
               32870      2      20      15     300                   1

   Port       Prio   Cost     State  Role Designated Bridge      Des Port Type
------------ ---- ---------- -----  ---- ---------------------- -------- ----
3      (pc2)   128      990!+  FWD   DESG 8066-08:17:f4:34:4d:00     801c  P2P
4      (pc2)   128      990!+  FWD   DESG 8066-08:17:f4:34:4d:00     801c  P2P
! = Automatic path cost.
+ = Portchannel cost, not the individual port cost.

----------------------------------------------------------------------
Spanning Tree Group 128: Off (PVRST), FDB aging timer 300
VLANs:  4095

   Port       Prio   Cost     State  Role Designated Bridge      Des Port Type
------------ ---- ---------- -----  ---- ---------------------- -------- ----
MGTA             0        0    FWD *
MGTB             0        0    FWD *
* = STP turned off for this port.
ACC-1#
```

In Example 5-32 on page 181, you can see that the current switch (ACC-1) is the root for VLAN 10 and the ACC-2 switch is the root for VLAN 20, by checking the MAC addresses of the root bridge for each STG.

You can find the MAC address of the current root by running **show system**:

```
ACC-1#show system | include MAC
MAC address: 08:17:f4:34:4d:00    IP (If 10) address: 10.0.10.2
MGMT-A Port MAC Address: 08:17:f4:34:4d:fe
MGMT-B Port MAC Address: 08:17:f4:34:4d:ef

ACC-2#show system | include MAC
MAC address: 08:17:f4:34:4c:00    IP (If 10) address: 10.0.10.3
MGMT-A Port MAC Address: 08:17:f4:34:4c:fe
MGMT-B Port MAC Address: 08:17:f4:34:4c:ef
ACC-2#
```

5.2.4 Quality of Service

A Quality of Service (QoS) implementation in IBM Networking OS is not in the scope of the reference architecture used in this book. However, a summary of configuration topics and commands is presented for completeness.

QoS commands configure the 802.1p priority value and DiffServ Code Point value of incoming packets, so you can differentiate between various types of traffic, and provide different priority levels.

802.1p configuration

This feature provides the switch the capability to filter IP packets based on the 802.1p bits in the packet's VLAN header. The 802.1p bits specify the priority that you should give to the packets while forwarding them. The packets with a higher (non-zero) priority bits are given forwarding preference over packets with numerically lower priority bits value.

You can use the following commands to complete the configuration:

- Run **qos transmit-queue mapping <priority (0-7)> <COSq number>** to map the 802.1p priority of to the Class of Service queue (COSq) priority. Enter the 802.1p priority value (0 - 7), followed by the Class of Service queue that handles the matching traffic.
- Run **qos transmit-queue weight-cos <COSq number> <weight (0-15)>** to configure the weight of the selected Class of Service queue (COSq). Enter the queue number (0 - 1), followed by the scheduling weight (0 - 15).
- Run **show qos transmit-queue** to show the current 802.1p parameters.

DSCP configuration

The following commands map the DiffServ Code Point (DSCP) value of incoming packets to a new value or to an 802.1p priority value:

- Run **qos dscp dscp-mapping <DSCP (0-63)> <new DSCP (0-63)>** to map the initial DiffServ Code Point (DSCP) value to a new value. Enter the DSCP value (0 - 63) of the incoming packets, followed by the new value.
- Run **qos dscp dot1p-mapping <DSCP (0-63)> <priority (0-7)>** to map the DiffServ Code point value to an 802.1p priority value. Enter the DSCP value, followed by the corresponding 802.1p value.

- Run **[no] qos dscp re-marking** to turn on/off DSCP re-marking globally.
- Run **show qos dscp** to display the current DSCP parameters.

Control plane protection

Important: This feature is supported only on RackSwitch G8264 switches.

You can use the following commands to limit the number of selected protocol packets received by the control plane (CP) of the switch. These limits help protect the CP from receiving too many protocol packets in a time period.

- Run **qos protocol-packet-control packet-queue-map <packet queue number (0-40)> <packet type>** to configure a packet type to associate with each packet queue number. Enter a queue number, followed by the packet type. You may map multiple packet types to a single queue. The following packet types are allowed:
  - 802.1x (IEEE 802.1x packets)
  - application-cri-packets (critical packets of various applications, such as Telnet and SSH)
  - arp-bcast (ARP broadcast packets)
  - arp-ucast (ARP unicast reply packets)
  - bgp (BGP packets)
  - bpdu (STP packets)
  - cisco-bpdu (Cisco STP packets)
  - dest-unknown (packets with destination not yet learned)
  - dhcp (DHCP packets)
  - icmp (ICMP packets)
  - igmp (IGMP packets)
  - ipv4-miscellaneous (IPv4 packets with IP options and TTL exception)
  - ipv6-nd (IPv6 Neighbor Discovery packets)
  - lacp (LACP/Link Aggregation protocol packets)
  - lldp (LLDP packets)
  - ospf (OSPF packets)
  - ospf3 (OSPF3 Packets)
  - pim (PIM packets)
  - rip (RIP packets)
  - system (system protocols, such as TFTP, FTP, Telnet, and SSH)
  - udld (UDLD packets)
  - vlag (VLAG packets)
  - vrrp (VRRP packets)
- Run **qos protocol-packet-control rate-limit-packetqueue <packet queue number (0-40)> <1-10000>** to configure the number of packets per second allowed for each packet queue.
- Run **no qos protocol-packet-control packet-queue-map <packet type>** to clear the selected packet type from its associated packet queue.

- Run **no qos protocol-packet-control rate-limit-packetqueue <packet queue number (0-40)>** to clear the packet rate configured for the selected packet queue.
- Run **show qos protocol-packet-control information protocol** to display the mapping of protocol packet types to each packet queue number. The status indicates whether the protocol is running or not running.
- Run **show qos protocol-packet-control information queue** to display the packet rate configured for each packet queue.

Weighted Random Early Detection configuration

Important: This feature is supported only on RackSwitch G8264 switches.

Weighted Random Early Detection (WRED) provides congestion avoidance by preemptively dropping packets before a queue becomes full. The RackSwitch G8264 implementation of WRED defines TCP and non-TCP traffic profiles on a per-port, per COS queue basis. For each port, you can define a transmit-queue profile with thresholds that define packet-drop probability.

To use WRED, run the following commands:

- Run **qos random-detect ecn** to enable or disable Explicit Congestion Notification (ECN). When ECN is on, the switch marks the ECN bit of the packet (if applicable) instead of dropping the packet. ECN-aware devices are notified of the congestion and those devices can take corrective actions.

  Important: ECN functions only on TCP traffic.

- Run **[no] qos random-detect enable** to turn on/off Random Detection and avoidance.
- Run **show qos random-detect** to display the current Random Detection and avoidance parameters.

WRED Transmit Queue configuration

Note: This feature is supported only on RackSwitch G8264.

To use WRED Transmit Queue, run the following commands:

- Run **[no] qos random detect transmit-queue <0-7> tcp <min. threshold (1-100)> <max.threshold (1-100)> <drop rate (1-100)>** to configure the WRED thresholds for TCP traffic. Use the **no** parameter to clear the WRED threshold value.
- Run **[no] qos random detect transmit-queue <0-7> non-tcp <min. threshold (1-100)> <max.threshold (1-100)> <drop rate (1-100)>** to configure the WRED thresholds for non-TCP traffic. Use the **no** parameter to clear the WRED threshold value.
- Run **[no] qos random detect transmit-queue <0-7> enable** to set the WRED transmit queue configuration to on or off.

5.3 Layer 3

This section provides configuration background information for using the RackSwitch G8124 and RackSwitch G8264 switches to perform IP routing functions. Any differences between the models regarding the availability of certain software features or command syntax are highlighted.

All the configurations presented in this chapter are implemented by using the IBM Networking OS V6.8software installed in the reference architecture switches. For planning details and the network diagrams used for implementation, see Chapter 3, “Reference architectures” on page 107

Not all the Layer 3 functions available in IBM Networking OS V6.8 are covered in this section. For an extensive list of features and configuration guidelines, see the documentation links listed in 5.5, “More information” on page 238.

The following topics are addressed in this chapter:

- Basic IPv4 configuration
- Basic IPv6 configuration
- OSPF configuration
- BGP configuration

5.3.1 Basic IPv4 configuration

The switches use a combination of configurable IP switch interfaces and IP routing options. The switches IP routing capabilities provide the following benefits:

- Connects the server IP subnets to the rest of the backbone network.
- Routes IP traffic between multiple VLANs configured on the switch.

This section covers configuration steps and commands for implementing basic routing between subnets using IP interfaces assigned to VLANs, default gateways, and static routes.

IP interfaces

The RackSwitch G8264 and RackSwitch G8124 switches support up to 128 IP interfaces. Each IP interface represents the switch on an IP subnet on the network. The Interface option is disabled by default.

In the RackSwitch G8264 switch, interface 128 is reserved for switch management on the MGT port.

In the RackSwitch G8124 switch, interface 127 and interface 128 are reserved for switch management, as follows:

- IF 127: Management port B
- IF 128: Management port A

To configure and enable an IP interface, complete the following steps:

1. Complete the VLAN selection.

   An IP interface must be mapped to a VLAN. Select one of the VLANs configured in the 5.2.1, “VLANs” on page 163 to apply an IP configuration to.

   VLAN 100 is used in this example. Perform the same configuration for all the VLANs, according to the input information in Chapter 3, “Reference architectures” on page 107.

2. Complete the IP interface configuration.

   Complete the following steps to create and activate an IP interface (use the commands shown in Example 5-33):

   a. Create the IP interface.

   b. Define an IP address.

   c. Map the IP interface to a VLAN.

   d. Enable the IP interface.

Example 5-33 IP interface configuration

```
AGG-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-1(config)#interface ip 100 (create the interface)
AGG-1(config-ip-if)#ip address 10.0.100.1 255.255.255.252 (define an IP address)
AGG-1(config-ip-if)#vlan 100 (map the IP interface to a VLAN)
AGG-1(config-ip-if)#enable (enable the IP interface)
AGG-1(config-ip-if)#^Z
AGG-1#
```

Interface and VLAN numbers: The IP interface number does not need to be the same as the VLAN number. The IP interfaces are 1 - 128, and the VLANs are 1 - 4095.

To keep the reference architecture simple and easy to understand, we used VLAN numbers lower than 128, which can be mapped to the same IP interface numbers.

The management interface is assigned by default to VLAN 4095 and also mapped to IP interface 128. Example 5-34 shows the management interface configuration.

Example 5-34 Management interface configuration

```
AGG-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-1(config)#interface ip 128
AGG-1(config-ip-if)#ip address 172.25.101.120
AGG-1(config-ip-if)#enable
AGG-1(config-ip-if)#^Z
AGG-1#
```

Run **no interface ip <*interface number*>** to remove an IP interface.

3. Verify the configuration.

   To see the basic IP configuration, run **show interface ip** (Example 5-35).

Example 5-35 IP configuration

```
AGG-1#show interface ip
Interface information:
100: IP4 10.0.100.1      255.255.255.252 10.0.100.3,      vlan 100, up
101: IP4 10.0.101.2      255.255.255.252 10.0.101.3,      vlan 101, up
103: IP4 10.0.103.1      255.255.255.252 10.0.103.3,      vlan 103, up
128: IP4 172.25.101.120  255.255.0.0     172.25.255.255,  vlan 4095, up
AGG-1#
```

Default gateway

Management IP addresses and gateways should already be configured during the initial setup. However, configuration guidelines are presented here for completeness.

Important: This feature is used only for management interfaces in our reference architecture. Static routes are not used in the reference architecture; instead, we use a dynamic routing protocol (OSPF).

The switch can be configured with up to four IPv4 gateways, as described in the following list:

RackSwitch G8124 Gateway 1 and Gateway 2: Data traffic

Gateway 3: Management port A

Gateway 4: Management port B

RackSwitch G8264 Gateway 1, Gateway 2, and Gateway 3: data traffic

Gateway 3: Management port

This gateway option is disabled by default. Complete the following steps to configure the default gateways:

1. Define the gateway.

   a. Configure the gateway by running **ip gateway <1-4> address <IP address>** as follows:

   ```
   AGG-1#configure terminal
   Enter configuration commands, one per line.  End with Ctrl/Z.
   AGG-1(config)#ip gateway 4 address 172.25.1.1
   AGG-1(config)#
   ```

   b. Configure additional parameters for the gateway health check, such as timers and number of tries. For more information about configuring these parameters, see 5.5, “More information” on page 238.

   If you need to, you can run **no ip gateway <1-4>** to delete the selected gateway.

2. Enable the gateway.

   Run **[no] ip gateway <1-4> enable** to enable or disable the gateway as follows:

   ```
   AGG-1#configure terminal
   Enter configuration commands, one per line.  End with Ctrl/Z.
   AGG-1(config)#ip gateway 4 enable
   AGG-1(config)#
   ```

3. Verify the gateway.

   Run **show ip gateway <1-4>** to display the gateway configuration parameters as follows:

   ```
   AGG-1#show ip gateway 4
   Current default gateway 4:
     172.25.1.1,      intr 2, retry 8, arp enabled, enabled
   AGG-1#
   ```

   Run **ping** to verify the gateway reachability, as shown in Example 5-36.

   Example 5-36 Ping gateway

   ```
   AGG-1#ping 172.25.1.1 mgt-port
   Connecting via MGT port.
   [host 172.25.1.1, max tries 5, delay 1000 msec , length 0]
   172.25.1.1: #1 ok, RTT 1 msec.
   ```

```
  172.25.1.1: #2 ok, RTT 1 msec.
  172.25.1.1: #3 ok, RTT 1 msec.
  172.25.1.1: #4 ok, RTT 1 msec.
  172.25.1.1: #5 ok, RTT 1 msec.
  Ping finished.
  AGG-1#
```

Important: The default **ping** destination is management port unless otherwise specified. To test IP addresses reachability through the data ports, use the **data-port** option with the **ping** command as follows:

```
AGG-1#ping 172.25.1.1 ?
  <0-4294967295>  Number of packets
  data-port       Data port
  mgt-port        Management port
  <cr>

AGG-1#
```

IPv4 static routes

Static routes are not used in the reference architecture. We use dynamic routing protocol (OSPF) instead. This section shows basic configuration commands and concepts for completeness. For more details, see 5.5, “More information” on page 238.

Static routes: Up to 128 IPv4 static routes can be configured only on RackSwitch G8264 and RackSwitch G8124 switches.

Run **ip route <IP subnet> <IP netmask> <IP nexthop> [<interface number>]** to define a static route (Example 5-37). Enter all the addresses by using dotted decimal notation.

Example 5-37 Static route configuration

```
ACC-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-1(config)#ip route 10.0.0.0 255.255.0.0 10.0.101.2
wait ...
ACC-1(config)#^Z
ACC-1#
```

Run **show ip route [static]** to display information about static routes (Example 5-38).

Example 5-38 Static route verification

```
ACC-1#sh ip route static
Current static routes:
  Destination     Mask            Gateway         If   ECMP    BGP
  --------------- --------------- --------------- ---- ------ ------
  10.0.0.0        255.255.0.0     10.0.101.2
ECMP health-check ping interval: 1
ECMP health-check retries number: 3
ECMP Hash Mechanism:  sip dip tcpl4 udpl4 sport dport
Gateway healthcheck: disabled
ACC-1#
```

Run **no ip route <IP subnet> <IP netmask> [<interface number>]** to remove a static route. The destination address of the route to remove must be specified by using dotted decimal notation.

Run **no ip route destination-address <IP address>** to clear all static routes with the specified destination.

Run **no ip route gateway <IP address>** to clear all static routes that use the specified gateway.

Equal-Cost Multi-Path static routes

Equal-Cost Multi-Path (ECMP) static routes are not used in the reference architecture. We use dynamic routing protocol (OSPF) instead. This section shows basic configuration commands and concepts for completeness. For more details, see 5.5, “More information” on page 238.

ECMP is a forwarding mechanism that routes packets along multiple paths of equal cost. ECMP provides equal-distributed link load sharing across the paths. The hashing algorithm used is based on the source IP address (SIP). ECMP routes allow the switch to choose between several next hops toward a destination. The switch performs periodic health checks (**ping**) on each ECMP gateway. If a gateway fails, it is removed from the routing table, and an SNMP trap is sent.

To configure ECMP static routes, add the same route multiple times, each with the same destination IP address, but with a different gateway IP address. These routes become ECMP routes. You can use two methods:

- OSPF integration: When a dynamic route is added through Open Shortest Path First (OSPF), the switch checks the route’s gateway against the ECMP static routes. If the gateway matches one of the single or ECMP static route destinations, then the OSPF route is added to the list of ECMP static routes. Traffic is load-balanced across all of the available gateways. When the OSPF dynamic route times out, it is deleted from the list of ECMP static routes.
- ECMP route hashing: You can configure the parameters used to perform ECMP route hashing as follows:
  - **sip**: Source IP address (default)
  - **dipsip**: Source IP address and destination IP address

sip and dipsip options: The **sip** and **dipsip** options enabled under ECMP route hashing or in port trunk hashing (portchannel hash) apply to both ECMP and trunk features (the enabled settings are cumulative). If unexpected ECMP route hashing occurs, disable the unwanted source or destination IP address option set in trunk hashing. Likewise, if unexpected trunk hashing occurs, disable any unwanted options set in ECMP route hashing.

Hash setting: The ECMP hash setting applies to all ECMP routes.

Use the following steps to configure ECMP static routes.

1. Define multiple static routes for the same destination but with different gateways.

   Run **ip route <IP subnet> <IP netmask> <IP nexthop> [<interface number>]** to define a static route. Enter all addresses by using dotted decimal notation, as shown in Example 5-39.

   Example 5-39 ECMP static routes configuration

```
ACC-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-1(config)#ip route 10.0.0.0 255.255.0.0 10.0.101.2
wait ...
ACC-1(config)#ip route 10.0.0.0 255.255.0.0 10.0.102.2
wait ...

Aug  9  8:59:03 ACC-1 NOTICE  ip: ECMP route configured, Gateway health check enabled
Aug  9  8:59:04 ACC-1 NOTICE  system: ECMP route gateway 10.0.101.2 is up
Aug  9  8:59:04 ACC-1 NOTICE  system: ECMP route gateway 10.0.102.2 is up

ACC-1(config)#^Z
ACC-1#
```

2. Configure hashing and health-check.

   Define additional ECMP static routes parameters such as hashing, health-check timers, and retries.

   Run **ip route ecmphash [sip][dip][protocol][tcpl4][udpl4][sport][dport]** to configure ECMP hashing. You may choose one or more of the following parameters:

   - **sip**: Source IP address (default)
   - **dip**: Destination IP address (default)
   - **protocol**: Layer 3 protocol
   - **tcpl4**: Layer 4 TCP traffic (default)
   - **udpl4**: Layer 4 UDP traffic (default)
   - **sport**: Source port (default)
   - **dport**: Destination port (default)

   Important: For G8264 switch, only sip and dip hashing parameters are available

   In Example 5-40, one more hashing parameter was configured in addition to the default settings.

   Example 5-40 Hashing parameter configuration

```
ACC-1#configure terminal
ACC-1(config)#ip route ecmphash protocol
Aug  9  9:26:00 ACC-1 WARNING cfg: ECMP hash got changed, Dataplane L3 hash
includes configured Trunk hash and ECMP hash
wait ...
ACC-1(config)#^Z
ACC-1#
```

   Run **ip route interval <1-60>** to configure the ECMP health-check **ping** interval, in seconds. The default value is 1 second.

   Run **ip route retries <1-60>** to configure the number of ECMP health-check retries. The default value is 3.

Run **[no] ip route healthcheck** to enable or disable static route health checks. The default setting is disabled.

3. Verify the ECMP static routes operation.

   Run **show ip route static** to display the static routes information. You can see the difference between simple static routes and ECMP static routes in Example 5-41.

Example 5-41 ECMP static routes verification

```
Simple static route
ACC-1#sh ip route static
Current static routes:
  Destination     Mask            Gateway         If   ECMP    BGP
  --------------- --------------- --------------- ---- ------ ------
  10.0.0.0        255.255.0.0     10.0.101.2
ECMP health-check ping interval: 1
ECMP health-check retries number: 3
ECMP Hash Mechanism:  sip dip tcpl4 udpl4 sport dport
Gateway healthcheck: disabled
ACC-1#

ECMP static route
ACC-1#show ip route static
Current static routes:
  Destination     Mask            Gateway         If   ECMP    BGP
  --------------- --------------- --------------- ---- ------ ------
  10.0.0.0        255.255.0.0     10.0.101.2           *
                                  10.0.102.2           *
ECMP health-check ping interval: 1
ECMP health-check retries number: 3
ECMP Hash Mechanism:  sip dip protocol tcpl4 udpl4 sport dport
Gateway healthcheck: enabled
ACC-1#
```

5.3.2 Basic IPv6 configuration

This section describes basic IPv6 implementation on the same network infrastructure as IPv4, with minimal configuration changes.

The goal is to provide a starting point for a seamless IPv4 to IPv6 migration and to show how IPv6 and IPv4 can coexist in the same network, by using the same hardware and software resources, with no additional investment and without interfering each other.

The IPv6 implementation for the reference architecture uses the same layout as IPv4.

Both protocols run on identical Layer 1 and Layer 2 layers. IPv6 uses a different set of IP interfaces, but has the same VLANs and STP configuration as IPv4. All the equipment is dual stacked and run IPv6 and IPv4 simultaneously (both the switches and the hosts). Servers are connected to switches using untagged ports, carrying IPv4 and IPv6 on the same VLAN.

Important: Carrying the IPv4 and the IPv6 traffic on the same VLAN facilitates an easy migration with minimal configuration changes on the network infrastructure.

The configuration is intended to be similar for both protocols, but some features are not available in the IBM Networking OS implementation of IPv6.

Important: For a complete list of the supported features, see the IBM Networking OS 6.8 Features Summary at:

http://www.ibm.com/support/docview.wss?uid=isg3T7000470

At the time of writing, the following IPv6 features were not supported in IBM Networking OS V6.8:

- Dynamic Host Control Protocol for IPv6 (DHCPv6)
- Border Gateway Protocol for IPv6 (BGP)
- Routing Information Protocol for IPv6 (RIPng)

You can use IBM Networking OS V6.8 features to configure IP addresses to use either IPv4 or IPv6 address formats. However, the following switch features support IPv4 only:

- SNMP trap host destination IP address
- Bootstrap Protocol (BOOTP) and DHCP
- RADIUS, TACACS+, and LDAP
- QoS metering and re-marking ACLs for out-profile traffic
- VMware Virtual Center (vCenter) for VMready
- Routing Information Protocol (RIP)
- Internet Group Management Protocol (IGMP)
- Border Gateway Protocol (BGP)
- Protocol Independent Multicast (PIM)
- Virtual Router Redundancy Protocol (VRRP)
- sFLOW

The main difference between the IPv6 and IPv4 implementation of the reference architecture described in Chapter 3, “Reference architectures” on page 107 is the restriction of using VRRP only for IPv4. There is no gateway redundancy protocol in IPv6 implementation. Anycast addresses are used in combination with server NIC teaming. The same gateway IPv6 address is configured as an anycast address on both access switches IP interfaces. Packets sent to an anycast address are delivered to the nearest interface identified by that address. The NIC teaming on the host is configured in *failover* mode, which means that only one port is active at a time (and connected to an access switch) and the other port is passive (and connected to the other access switch). Using NIC teaming, packets are sent only to an access switch.

For the implementation input information (IP addresses, VLANs, trunks, and so on), see Chapter 3, “Reference architectures” on page 107.

IPv6 interfaces

The following guidelines apply for IPv6 IP interfaces on the RackSwitch G8264 and RackSwitch G8124 switches:

- Each IPv6 interface supports multiple IPv6 addresses. You can manually configure up to two IPv6 addresses for each interface, or you can allow the switch to use stateless autoconfiguration.
- You can manually configure two IPv6 addresses for each interface, as follows:
  - The initial IPv6 address is a global unicast or anycast address.
  - A second IPv6 address can be a unicast or anycast address.
  - You cannot configure both addresses as anycast. If you configure an anycast address on the interface, you must also configure a global unicast address on that interface.

- You cannot configure an IPv4 address on an IPv6 interface. Each interface can be configured with only one address type, either IPv4 or IPv6, but not both. When changing between IPv4 and IPv6 address formats, the prior address settings for the interface are discarded.
- Each IPv6 interface can belong to only one VLAN.
- Each VLAN can support only one IPv6 interface.
- Each VLAN can support multiple IPv4 interfaces.

IPv4 layout for IPv6: An easy method of using the IPv4 layout for the IPv6 implementation is to configure separate IPv6 interfaces, but to map them to the same VLANs used by IPv4, thus having two IP interfaces that use a single VLAN.

Each IPv6 interface can be configured as a router node or a host node, as follows:

- A router node's IP address is configured manually. Router nodes can send Router Advertisements.
- A host node's IP address is autoconfigured. The host nodes listen for Router Advertisements that convey information about devices on the network. You can manually assign an IPv6 address to an interface in host mode, or the interface can be assigned an IPv6 address by an upstream router, using information from router advertisements to perform stateless auto-configuration.

IPv6 interfaces:

- All the IPv6 interfaces on point-to-point links, between aggregation and access switches, used for OSPF routing, are configured as *host nodes* with static IPv6 addresses.
- All the IPv6 interfaces connected to servers are configured as *router nodes*, with Router Advertisement enabled so the listening hosts can perform stateless auto-configuration.

Complete the following steps to configure IPv6 interfaces:

1. Configure the host nodes interfaces.

   To configure host nodes interfaces, run **`ip6host`** and **`ipv6 address <IPv6 address> <prefix length> [anycast|enable]`** in the interface configuration mode (Example 5-42).

   Example 5-42 Host node interface configuration

```
AGG-1#conf t
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-1(config)#interface ip 111
AGG-1(config-ip-if)#ip6host
AGG-1(config-ip-if)#ipv6 address fc11::2 64
AGG-1(config-ip-if)#vlan 101
AGG-1(config-ip-if)#enable
AGG-1(config-ip-if)#^Z
AGG-1#
```

   IPv6 address: If no IPv6 address is specified, the host node is listening for Router Advertisements for auto-configuration information.

   Configure all the interfaces between aggregation and access switches as host nodes, according to the input information in Chapter 3, "Reference architectures" on page 107.

IPv6 interfaces and VLAN: The IPv6 interfaces are mapped to the same VLAN as their IPv4 pairs:

```
AGG-1#
interface ip 101
        ip address 10.0.101.2 255.255.255.252
        vlan 101
        enable
        exit

interface ip 111
        ipv6 address fc11:0:0:0:0:0:0:2 64
        vlan 101
        enable
        ip6host
        exit
AGG-1#
```

2. Configure the router nodes interfaces (Example 5-43).

   Run **ipv6 address <IPv6 address> <prefix length> [anycast|enable]** to configure a primary unicast IPv6 address on an interface.

   Run **ipv6 secaddr6 address <IPv6 address> <prefix length> [anycast]** to assign a secondary anycast address to an interface.

Example 5-43 Router node interface configuration

```
ACC-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-1(config)#interface ip 106
ACC-1(config-ip-if)#ipv6 address fc10::2 64
ACC-1(config-ip-if)#ipv6 secaddr6 address fc10::1 64 anycast
ACC-1(config-ip-if)#vlan 10
ACC-1(config-ip-if)#enable
ACC-1(config-ip-if)#^Z
ACC-1#
```

Aggregation and access switches: Configure all the interfaces between aggregation and access switches as router nodes, using the input information from Chapter 3, “Reference architectures” on page 107.

We use static IPv6 address and default gateway assignment for the hosts in our architecture, but for completeness, we provide the basic Router Advertisements configuration. For detailed information, see 5.5, “More information” on page 238.

To enable Router Advertisements for the specified interface, run the **ipv6 nd** related commands.

Run **[no] ipv6 nd suppress-ra** to enable or disable IPv6 Router Advertisements on the interface. The default setting is disabled (suppress Router Advertisements), so use the **no** parameter to activate them.

Run **ipv6 nd prefix <IPv6 address> <prefix length>** to define the prefix to be advertised (Example 5-44).

Example 5-44 Neighbor Discover protocol configuration

```
ACC-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-1(config)#interface ip 106
ACC-1(config-ip-if)#no ipv6 nd suppress-ra
ACC-1(config-ip-if)#ipv6 nd prefix fc10:0:0:0:0:0:0:0 64
ACC-1(config-ip-if)#^Z
ACC-1#
```

3. Host configuration.

   We now define the host part of the configuration for Windows and Linux.

   – Configure Windows hosts.

     For the Top-of-Rack reference architecture, we used an IBM System x3550 M3 machine with a Windows Server 2008 R2 operating system connected to the two G8124 access switches (ACC1 and ACC2). For more details, see Chapter 3, “Reference architectures” on page 107.

     The server is connected to the switches by using a dual port Emulex 10GbE Virtual Fabric Adapter NIC, with vNIC mode disabled.

     The two network ports are configured in Failover mode teaming. To see the teaming configuration on the server, see Figure 5-2.

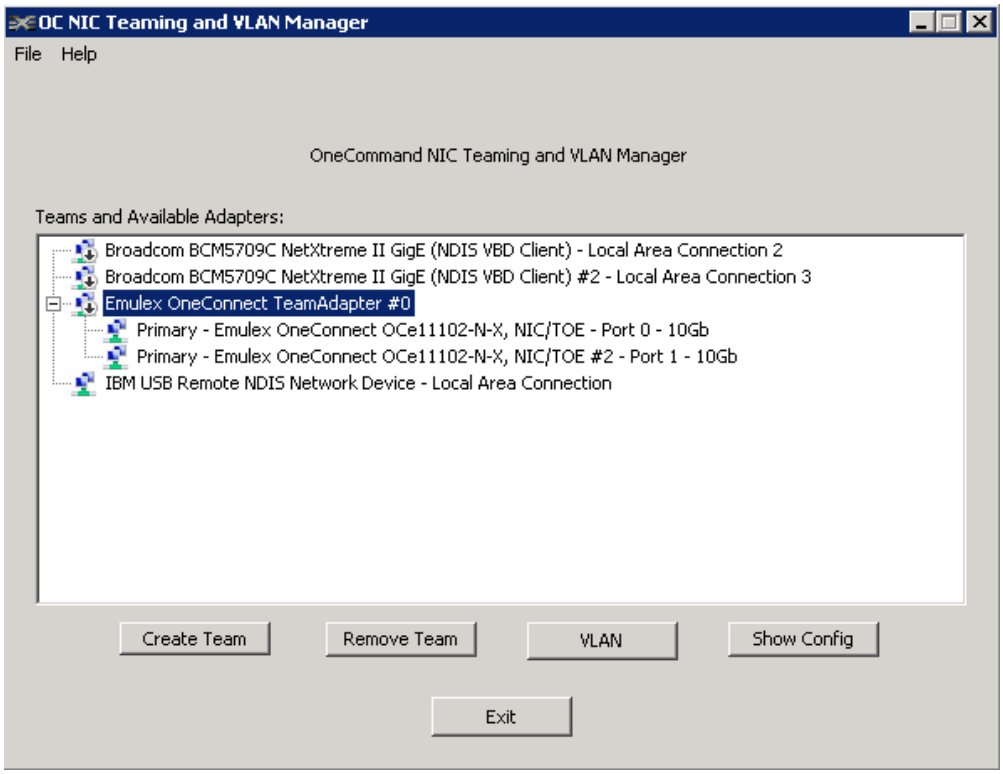
OC NIC Teaming and VLAN Manager
File
Help
OneCommand NIC Teaming and VLAN Manager
Teams and Available Adapters:
Broadcom BCM5709C NetXtreme II GigE (NDIS VBD Client) - Local Area Connection 2
Broadcom BCM5709C NetXtreme II GigE (NDIS VBD Client) #2 - Local Area Connection 3
Emulex OneConnect TeamAdapter #0
Primary - Emulex OneConnect OCe11102-N-X, NIC/TOE - Port 0 - 10Gb
Primary - Emulex OneConnect OCe11102-N-X, NIC/TOE #2 - Port 1 - 10Gb
IBM USB Remote NDIS Network Device - Local Area Connection
Create Team
Remove Team
VLAN
Show Config
Exit

Figure 5-2 NIC Teaming on server host

IPv4 and IPv6 are configured on the same server interface (Teaming Adapter). There is no need for additional configuration of the switch port, as shown in Example 5-45.

Example 5-45 Switch port configuration for a dual TCP/IP stack host

```
ACC-1#show interface information 7
Alias   Port Tag RMON Lrn Fld PVID       NAME                 VLAN(s)
------- ---- --- ---- --- --- ----- --------------
-------------------------------
7       7     n   d    e   e    10  SRV1           10

* = PVID is tagged.
ACC-1#

ACC-1#show running-config | begin 7
interface port 7
        name "SRV1"
        pvid 10
        exit
ACC-1#
```

The IPv4 and IPv6 configuration is implemented on the TeamAdapter interface that aggregates the two 10 Gb ports. The operating system is dual stack enabled, as shown in Figure 5-3.

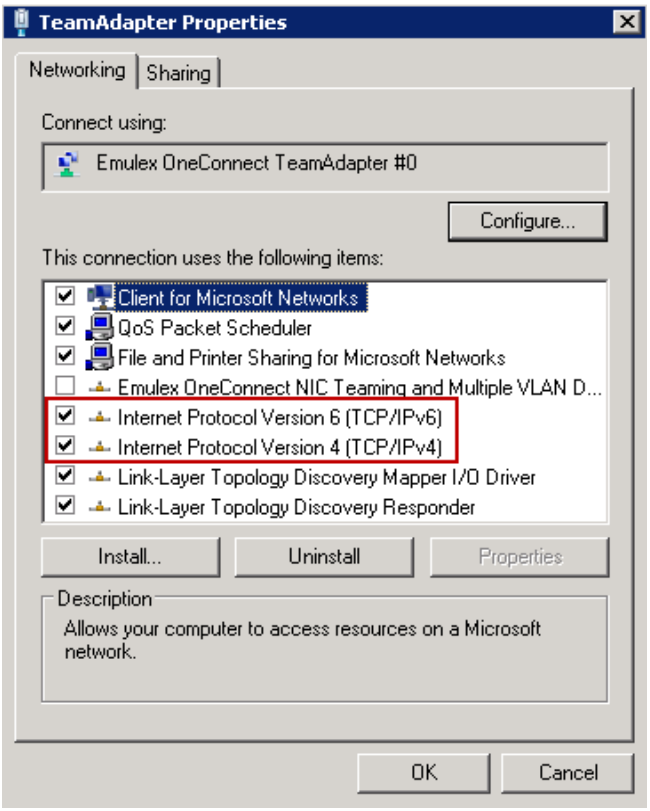

Figure 5-3 Dual TCP/IP stack in Windows

IPv4 and IPv6 addresses are configured by using the input information from Chapter 3, “Reference architectures” on page 107. The static IPv4 and IPv6 configuration on the interface is shown in Figure 5-4.

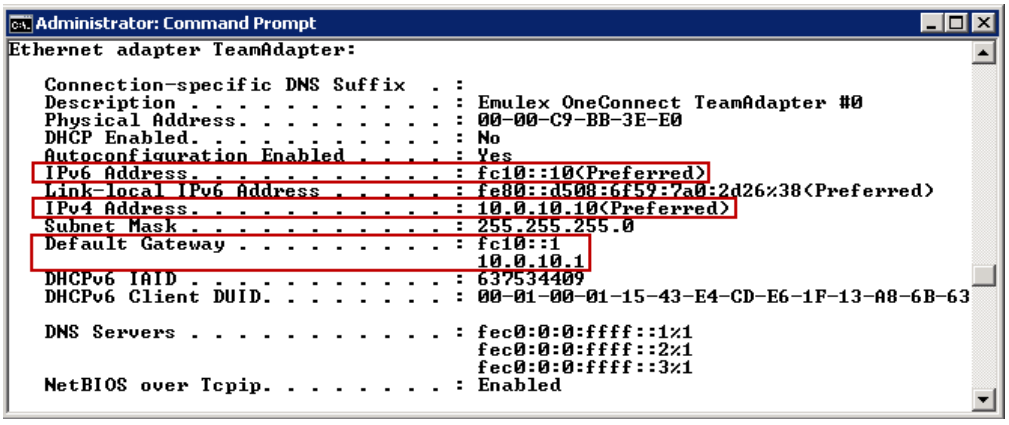
Administrator: Command Prompt
Ethernet adapter TeamAdapter:
Connection-specific DNS Suffix . :
Description . . . . . : Emulex OneConnect TeamAdapter #0
Physical Address. . . . . : 00-00-C9-BB-3E-E0
DHCP Enabled. . . . . : No
Autoconfiguration Enabled . . . . : Yes
IPv6 Address. . . . . : fc10::10(Preferred)
Link-local IPv6 Address . . . . . : fe80::d508:6f59:7a0:2d26%38(Preferred)
IPv4 Address. . . . . : 10.0.10.10(Preferred)
Subnet Mask . . . . . : 255.255.255.0
Default Gateway . . . . . : fc10::1
10.0.10.1
DHCPv6 IAID . . . . . : 637534409
DHCPv6 Client DUID. . . . . : 00-01-00-01-15-43-E4-CD-E6-1F-13-A8-6B-63
DNS Servers . . . . . : fec0:0:0:ffff::1%1
fec0:0:0:ffff::2%1
fec0:0:0:ffff::3%1
NetBIOS over Tcpip. . . . . : Enabled

Figure 5-4 Windows IPv4 and IPv6 static addresses configuration

If the Router Advertisements protocol is enabled on the switch (see step 2 on page 195), then the IPv6 interface on the host auto-configures itself using the advertised prefix, as shown in Figure 5-5.

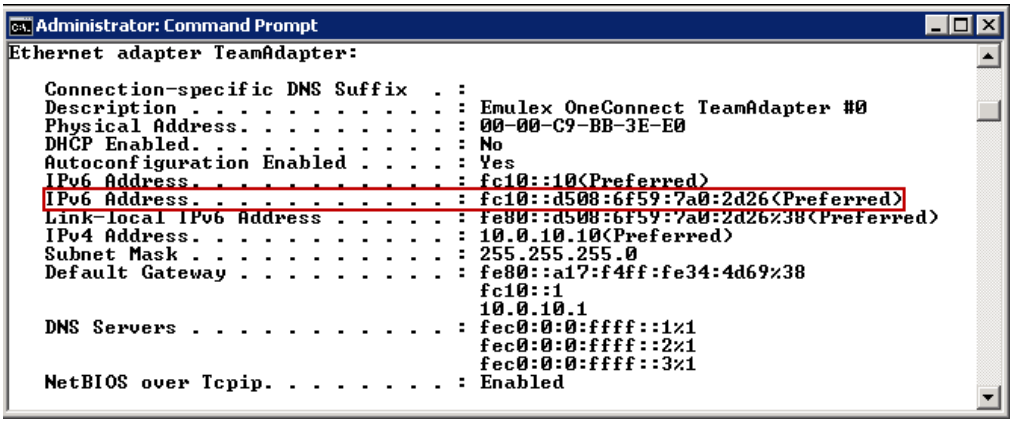
Administrator: Command Prompt
Ethernet adapter TeamAdapter:
Connection-specific DNS Suffix . :
Description . . . . . : Emulex OneConnect TeamAdapter #0
Physical Address. . . . . : 00-00-C9-BB-3E-E0
DHCP Enabled. . . . . : No
Autoconfiguration Enabled . . . . : Yes
IPv6 Address. . . . . : fc10::10(Preferred)
IPv6 Address. . . . . : fc10::d508:6f59:7a0:2d26(Preferred)
Link-local IPv6 Address . . . . . : fe80::d508:6f59:7a0:2d26%38(Preferred)
IPv4 Address. . . . . : 10.0.10.10(Preferred)
Subnet Mask . . . . . : 255.255.255.0
Default Gateway . . . . . : fe80::a17:f4ff:fe34:4d69%38
fc10::1
10.0.10.1
DNS Servers . . . . . : fec0:0:0:ffff::1%1
fec0:0:0:ffff::2%1
fec0:0:0:ffff::3%1
NetBIOS over Tcpip. . . . . : Enabled

Figure 5-5 Windows IPv6 auto-configuration

Both default gateways (IPv4 and IPv6) are reachable, as shown in Figure 5-6.

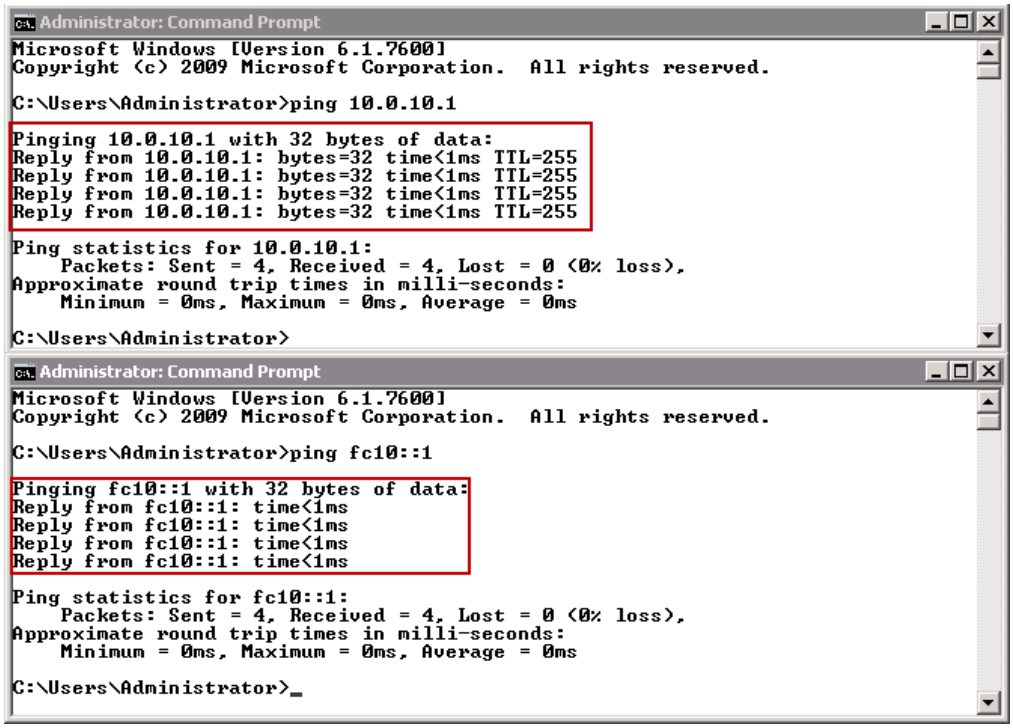

Figure 5-6 IPv4 and IPv6 gateway reachability

– Configure Linux hosts.

For the Embedded switch reference architecture (see Chapter 6, “IBM Virtual Fabric 10Gb Switch Module implementation” on page 239), we used a blade server that runs the Red Hat Enterprise Linux operating system connected internally to the two 10GbE Virtual Fabric Switches in the Blade Center chassis.

The blade server is installed in slot 14 and has internal ports 1:14 and 2:14 in the embedded switch.

Both 10 GbE interfaces of the server are configured in active-backup bonding. The IP configuration is applied on the bond1 interface, aggregating the two 10 Gb interfaces, numbered eth2 and eth3 in the operating system.

For the embedded switch architecture, VLAN30 is used for carrying both IPv4 and IPv6 traffic. There is no need for additional configuration from the IPv4 section. The switch configuration is the same as for Top-of-Rack, and is shown in Example 5-46.

Example 5-46 Switch port configuration for TCP/IP dual stack Linux host

```
ACC-3#sh interface information 1:14,2:14
Alias   Port Tag    Type    RMON Lrn Fld PVID       NAME
VLAN(s)
------- ---- --- ---------- ---- --- --- ----- -------------
------------------------------
1:14    14    y  Internal    d    e   e    30  SRV-3          30
2:14    78    y  Internal    d    e   e    30  SRV-3          30

* = PVID is tagged.
ACC-3#
```

```
ACC-3#show running-config | begin 1:14
interface port 1:14
        name "SRV-3"
        pvid 30
        exit

ACC-3#show running-config | begin 2:14
interface port 2:14
        name "SRV-3"
        pvid 30
        exit
ACC-3#
```

We start with the bonding interface already created but with no IP address, as shown in Figure 5-7.

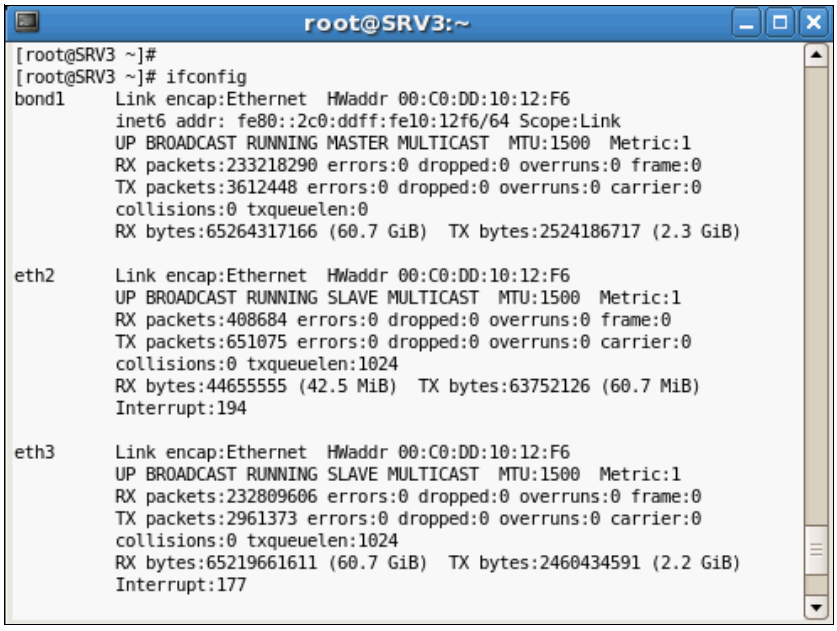
root@SRV3:~
[root@SRV3 ~]#
[root@SRV3 ~]# ifconfig
bond1 Link encap:Ethernet HWaddr 00:C0:DD:10:12:F6
inet6 addr: fe80::2c0:ddff:fe10:12f6/64 Scope:Link
UP BROADCAST RUNNING MASTER MULTICAST MTU:1500 Metric:1
RX packets:233218290 errors:0 dropped:0 overruns:0 frame:0
TX packets:3612448 errors:0 dropped:0 overruns:0 carrier:0
collisions:0 txqueuelen:0
RX bytes:65264317166 (60.7 GiB) TX bytes:2524186717 (2.3 GiB)
eth2 Link encap:Ethernet HWaddr 00:C0:DD:10:12:F6
UP BROADCAST RUNNING SLAVE MULTICAST MTU:1500 Metric:1
RX packets:408684 errors:0 dropped:0 overruns:0 frame:0
TX packets:651075 errors:0 dropped:0 overruns:0 carrier:0
collisions:0 txqueuelen:1024
RX bytes:44655555 (42.5 MiB) TX bytes:63752126 (60.7 MiB)
Interrupt:194
eth3 Link encap:Ethernet HWaddr 00:C0:DD:10:12:F6
UP BROADCAST RUNNING SLAVE MULTICAST MTU:1500 Metric:1
RX packets:232809606 errors:0 dropped:0 overruns:0 frame:0
TX packets:2961373 errors:0 dropped:0 overruns:0 carrier:0
collisions:0 txqueuelen:1024
RX bytes:65219661611 (60.7 GiB) TX bytes:2460434591 (2.2 GiB)
Interrupt:177

Figure 5-7 Red Hat Linux host network interfaces initial configuration

Because the Router Advertisement protocol is enabled on the switch, the IPv6 address is auto-configured on the bond interface, as shown in Figure 5-8.

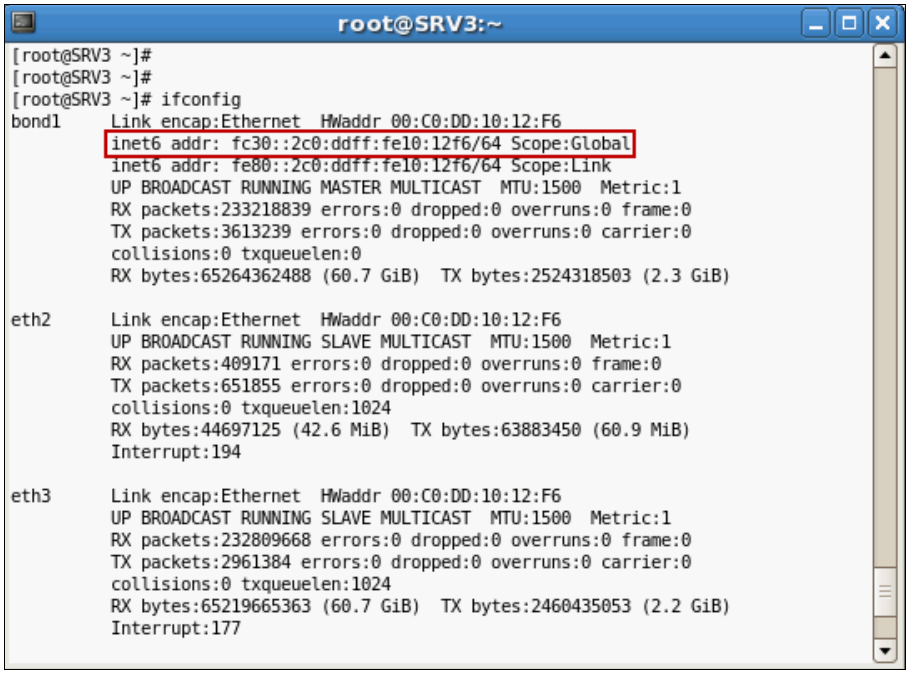
root@SRV3:~
[root@SRV3 ~]#
[root@SRV3 ~]#
[root@SRV3 ~]# ifconfig
bond1 Link encap:Ethernet HWaddr 00:C0:DD:10:12:F6
inet6 addr: fc30::2c0:ddff:fe10:12f6/64 Scope:Global
inet6 addr: fe80::2c0:ddff:fe10:12f6/64 Scope:Link
UP BROADCAST RUNNING MASTER MULTICAST MTU:1500 Metric:1
RX packets:233218839 errors:0 dropped:0 overruns:0 frame:0
TX packets:3613239 errors:0 dropped:0 overruns:0 carrier:0
collisions:0 txqueuelen:0
RX bytes:65264362488 (60.7 GiB) TX bytes:2524318503 (2.3 GiB)
eth2 Link encap:Ethernet HWaddr 00:C0:DD:10:12:F6
UP BROADCAST RUNNING SLAVE MULTICAST MTU:1500 Metric:1
RX packets:409171 errors:0 dropped:0 overruns:0 frame:0
TX packets:651855 errors:0 dropped:0 overruns:0 carrier:0
collisions:0 txqueuelen:1024
RX bytes:44697125 (42.6 MiB) TX bytes:63883450 (60.9 MiB)
Interrupt:194
eth3 Link encap:Ethernet HWaddr 00:C0:DD:10:12:F6
UP BROADCAST RUNNING SLAVE MULTICAST MTU:1500 Metric:1
RX packets:232809668 errors:0 dropped:0 overruns:0 frame:0
TX packets:2961384 errors:0 dropped:0 overruns:0 carrier:0
collisions:0 txqueuelen:1024
RX bytes:65219665363 (60.7 GiB) TX bytes:2460435053 (2.2 GiB)
Interrupt:177

Figure 5-8 Linux host IPv6 autoconfiguration

For our reference architecture, we use static IPv6 configuration. For a static IPv4 and IPv6 configuration example, see Figure 5-9.

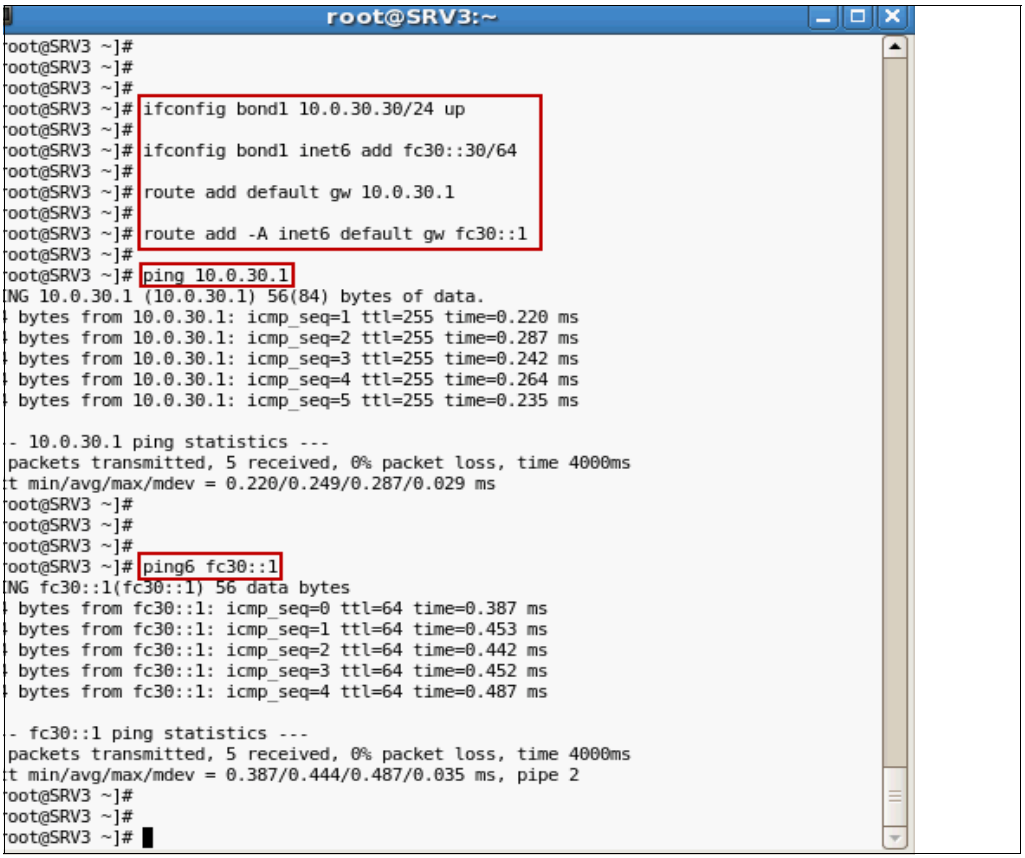
root@SRV3:~
root@SRV3 ~]#
root@SRV3 ~]#
root@SRV3 ~]#
root@SRV3 ~]# ifconfig bond1 10.0.30.30/24 up
root@SRV3 ~]#
root@SRV3 ~]# ifconfig bond1 inet6 add fc30::30/64
root@SRV3 ~]#
root@SRV3 ~]# route add default gw 10.0.30.1
root@SRV3 ~]#
root@SRV3 ~]# route add -A inet6 default gw fc30::1
root@SRV3 ~]#
root@SRV3 ~]# ping 10.0.30.1
NG 10.0.30.1 (10.0.30.1) 56(84) bytes of data.
bytes from 10.0.30.1: icmp_seq=1 ttl=255 time=0.220 ms
bytes from 10.0.30.1: icmp_seq=2 ttl=255 time=0.287 ms
bytes from 10.0.30.1: icmp_seq=3 ttl=255 time=0.242 ms
bytes from 10.0.30.1: icmp_seq=4 ttl=255 time=0.264 ms
bytes from 10.0.30.1: icmp_seq=5 ttl=255 time=0.235 ms
- 10.0.30.1 ping statistics ---
packets transmitted, 5 received, 0% packet loss, time 4000ms
t min/avg/max/mdev = 0.220/0.249/0.287/0.029 ms
root@SRV3 ~]#
root@SRV3 ~]#
root@SRV3 ~]#
root@SRV3 ~]# ping6 fc30::1
NG fc30::1(fc30::1) 56 data bytes
bytes from fc30::1: icmp_seq=0 ttl=64 time=0.387 ms
bytes from fc30::1: icmp_seq=1 ttl=64 time=0.453 ms
bytes from fc30::1: icmp_seq=2 ttl=64 time=0.442 ms
bytes from fc30::1: icmp_seq=3 ttl=64 time=0.452 ms
bytes from fc30::1: icmp_seq=4 ttl=64 time=0.487 ms
- fc30::1 ping statistics ---
packets transmitted, 5 received, 0% packet loss, time 4000ms
t min/avg/max/mdev = 0.387/0.444/0.487/0.035 ms, pipe 2
root@SRV3 ~]#
root@SRV3 ~]#
root@SRV3 ~]#

Figure 5-9 Linux static IPv4 and IPv6 configuration

Important: This method is the only method for configuring IP addresses and IP routes on Linux. This configuration is also a temporary configuration that is discarded at reboot. For a permanent IP configuration in Linux, see the appropriate Red Hat Linux documentation.

Open Shortest Path First

This section describes OSPF dynamic routing protocol implementation in switches using the reference architecture in Chapter 3, “Reference architectures” on page 107. It does not provide a complex configuration, as you would expect in an enterprise or service provider production network. It shows basic configuration steps only, as a starting point for enabling OSPF-based dynamic routing in the network for both IPv4 and IPv6 protocols, using IBM System Networking RackSwitches switches running IBM Networking OS.

Not all the features of OSPF protocol implemented in IBM Networking OS are covered in this section. For more detailed configuration options and commands, see 5.5, “More information” on page 238.

Our goal is to have a basic dynamic routing mechanism that runs on IPv4 and IPv6 at the same time, using the same infrastructure, without a significant configuration effort. For OSPFv3, you need IPv6 addresses, but restrictions apply when configuring IPv4 and IPv6 interfaces in IBM Networking OS.

Note: The following guidelines and limitations apply when configuring IPv4 and IPv6 interfaces:

- You cannot configure an IPv4 address on an IPv6 interface. Each interface can be configured with only one address type, either IPv4 or IPv6, but not both.
- Each IPv6 interface can belong to only one VLAN.
- Each VLAN can support only one IPv6 interface.
- Each VLAN can support multiple IPv4 interfaces.

The solution for using the same layout for both IP protocols is to have pairs of IPv4 and IPv6 interfaces mapped to the same VLAN.

Figure 5-10 shows the reference architecture diagram used for OSPF implementation. The two servers are Layer 2 connected to the core network. AGG-1, AGG-2, ACC-1, and ACC-2 are interconnected with point-to-point Layer 3 links. No VLAN spans the network.

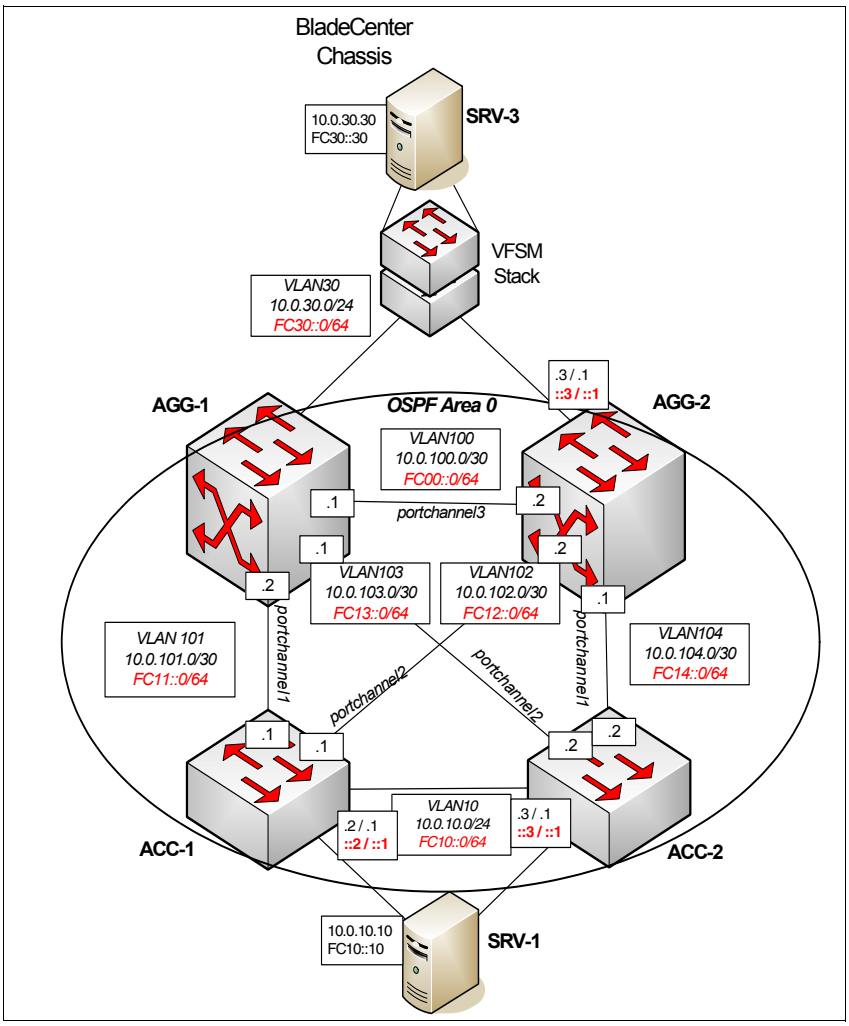

Figure 5-10 Network diagram used for OSPF implementation

For SRV-1 to reach SRV-3 on the other side of the network, it must rely on the routing mechanisms in the middle. OSPF must redistribute the directly connected networks of SRV-1 and SRV-3 to the other routers on the network.

The task is considered complete when SRV-1 is able to reach SRV-3 on both IPv4 and IPv6.

Reachability: The reachability test between SRV-1 and SRV-3 assumes that the 10Gb Virtual Fabric Switch configuration is complete and operational. See Chapter 6, “IBM Virtual Fabric 10Gb Switch Module implementation” on page 239.

OSPFv2 configuration

The OSPF configuration starts from the assumption that all the configuration tasks presented so far in this chapter have been performed. All the ports are up and running, all the VLANs are correctly defined, all the IP interfaces are configured and the routers are reachable on the Layer 3 links, and the servers are able to **`ping`** the default gateway.

The goal of the configuration is to enable OSPF, to configure the Layer 3 links in the protocol, and to redistribute the connected networks (server subnet) to the neighbors.

The final result is that a route to SRV-1 subnet (10.0.10.0/24) is installed in AGG-1 and AGG-2 routing table, a route to SRV-3 subnet (10.0.30.0/24) is installed in ACC-1 and ACC-2 routing table, and the SRV-1 and SRV-3 are able to **`ping`** each other.

Complete the following steps to perform a basic OSPF configuration. For more details and options, see 5.5, “More information” on page 238.

1. Define a Router ID.

   Routing devices in OSPF areas are identified by a router ID. The router ID is expressed in IP address format. The IP address of the router ID is not required to be included in any IP interface range or in any OSPF area.

   The router ID can be configured in one of the following two ways:

   Dynamically The OSPF protocol configures the lowest IP interface IP address as the router ID (this way is the default).

   Statically Run **`ip router-id <IPv4 address>`** to manually configure the router ID.

   To modify the router ID from static to dynamic, set the router ID to 0.0.0.0, save the configuration, and reboot the switch.

   For our reference architecture, an IP loopback interface is configured for each switch running OSPF, as shown in Example 5-47.

Example 5-47 Router ID and loopback interface configuration

```
AGG-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-1(config)#interface loopback 1
AGG-1(config-ip-loopback)#ip address 1.1.1.1 255.255.255.255
AGG-1(config-ip-loopback)#ip router-id 1.1.1.1
AGG-1(config)#^Z
AGG-1#


AGG-2#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-2(config)#interface loopback 1
```

```
AGG-2(config-ip-loopback)#ip address 1.1.1.2 255.255.255.255
AGG-2(config-ip-loopback)#ip router-id 1.1.1.2
AGG-2(config)#^Z
AGG-2#


ACC-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-1(config)#interface loopback 1
ACC-1(config-ip-loopback)#ip address 2.2.2.1 255.255.255.255
ACC-1(config-ip-loopback)#ip router-id 2.2.2.1
ACC-1(config)#^Z
ACC-1#


ACC-2#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-2(config)#interface loopback 1
ACC-2(config-ip-loopback)#ip address 2.2.2.2 255.255.255.255
ACC-2(config-ip-loopback)#ip router-id 2.2.2.2
ACC-2(config-ip-loopback)#^Z
ACC-2#
```

2. Enable OSPF.

   Run **`router ospf`** to enter the protocol configuration mode and **`enable`** to activate it, as shown in Example 5-48.

   Example 5-48 OSPF activation

```
ACC-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-1(config)#router ospf
ACC-1(config-router-ospf)#enable
ACC-1(config-router-ospf)#^Z
ACC-1#
```

   Important: This operation is performed for all four Layer 3 switches (AGG-1, AGG-2, ACC-1, and ACC-2)

3. Define areas.

   If you are configuring multiple areas in your OSPF domain, one of the areas must be designated as area 0, known as the *backbone*. The backbone is the central OSPF area and is physically connected to all other areas. The areas inject routing information into the backbone, which in turn disseminates the information into other areas.

   We use only area 0 (backbone) in our example.

   Up to six OSPF areas can be connected to the switch with IBM Networking OS. To configure an area, the OSPF number must be defined and then attached to a network interface on the switch. The full process is explained in the following sections.

   An OSPF area is defined by assigning two pieces of information: an *area index* and an *area ID*. The commands to define and enable an OSPF area are:

   - **`area <area index> area-id <n.n.n.n>`**
   - **`area <area index> enable`**

Important: The **area** option is an arbitrary index used only on the switch and does not represent the actual OSPF area number. The actual OSPF area number is defined in the **area-id**.

The **area <area index>** option is an arbitrary index (0 - 5) used only by the switch. This index number does not necessarily represent the OSPF area number, though for configuration simplicity, it should do so where possible. The area index and area ID configuration is shown in Example 5-49.

Example 5-49 Area index and area-id configuration

```
AGG-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-1(config)#router ospf
AGG-1(config-router-ospf)#area 0 area-id 0.0.0.0
AGG-1(config-router-ospf)#area 0 enable
AGG-1(config-router-ospf)#^Z
AGG-1#
```

For more details about **area-id** formats and options, see 5.5, “More information” on page 238.

Important: This operation is performed for all four Layer 3 switches (AGG-1, AGG-2, ACC-1, and ACC-2)

4. Configure authentication (optional).

   OSPF protocol exchanges can be authenticated so that only trusted routing devices can participate. This situation ensures less processing on routing devices that are not listening to OSPF packets. OSPF allows packet authentication and uses IP multicast when sending and receiving packets. Routers participate in routing domains based on predefined passwords. IBM Networking OS supports simple password (type 1 plain text passwords) and MD5 cryptographic authentication. This type of authentication allows a password to be configured per area (Example 5-50).

   Run **area <index> authentication-type {md5|password}** to configure the authentication type.

   We use the MD5 method, so a key must be defined by running **message-digest-key <index> md5-key <key>**.

   Example 5-50 Authentication configuration

```
AGG-1(config)#router ospf
AGG-1(config-router-ospf)#area 0 authentication-type md5
AGG-1(config-router-ospf)#message-digest-key 1 md5-key redbk
AGG-1(config-router-ospf)#^Z
AGG-1#
```

   Important: This operation is performed for all four Layer 3 switches (AGG-1, AGG-2, ACC-1, and ACC-2)

5. Attach an area to a network.

   After an OSPF area is defined, it must be associated with a network. To attach the area to a network, you must assign the OSPF area index to an IP interface that participates in the area. Run **ip ospf area <index>** and **ip ospf enable** under the interface configuration mode, as shown in Example 5-51.

   If authentication with MD5 is used, assign an MD5 key ID to OSPF interfaces by running **ip ospf message-digest-key <index>**.

   Example 5-51 Attach an area to a network

```
AGG-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-1(config)#interface ip 100
AGG-1(config-ip-if)#ip ospf area 0
AGG-1(config-ip-if)#ip ospf enable
AGG-1(config-ip-if)#ip ospf message-digest-key 1
AGG-1(config-ip-if)#^Z
AGG-1#
```

   Important: This operation is performed for all four Layer 3 switches (AGG-1, AGG-2, ACC-1, and ACC-2) for the IP interfaces 100, 101, 102, 103, and 104.

6. Configure route redistribution.

   We use basic redistribution of directly connected (fixed) networks in OSP. For more details and advanced configuration, see 5.5, “More information” on page 238.

   Run **redistribute {ebgp|fixed|ibgp|rip|static} {export|route_map} <metric> <AS external metric type>** to redistribute routes. We do not use route maps, but a simple export. Example 5-52 shows the fixed routes redistribution.

   Use equal metric for AGG-1 and ACC-1 and higher metrics for AGG-2 and ACC-2 because we want the traffic from SRV-1 to SRV-3 to go through AGG-1 and ACC-1.

   Example 5-52 Fixed routes redistribution

```
AGG-1#conf t
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-1(config)#router ospf
AGG-1(config-router-ospf)#redistribute fixed export 2 1
AGG-1(config-router-ospf)#^Z
AGG-1#

ACC-1#conf t
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-1(config)#router ospf
ACC-1(config-router-ospf)#redistribute fixed export 2 1
ACC-1(config-router-ospf)#^Z
ACC-1#

AGG-2#conf t
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-2(config)#router ospf
AGG-2(config-router-ospf)#redistribute fixed export 5 1
AGG-2(config-router-ospf)#^Z
AGG-2#
```

```
ACC-2#conf t
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-2(config)#router ospf
ACC-2(config-router-ospf)#redistribute fixed export 10 1
ACC-2(config-router-ospf)#^Z
ACC-2#
```

7. Verify OSPF.

   Run the following commands to verify OSPF operation:

   - **show ip ospf**
   - **show ip ospf neighbor**
   - **show ip ospf database database**
   - **show ip ospf routes**
   - **show ip route**

   See the following examples from ACC-1 switch for these commands' output.

   Show the current OSPF configuration settings by running **show ip ospf** (Example 5-53).

Example 5-53 show ip ospf command output

```
ACC-1#show ip ospf
Current OSPF settings: ON
  Default route none
  Router ID: 2.2.2.1
  lsdb limit 0

Current OSPF area settings:
  0: 0.0.0.0,         type transit, auth md5, metric 1, spf 10, enabled

Current OSPF interface settings:
101: 10.0.101.1,      area index  0, prio 1, cost 1, enabled
    hello 10, dead 40, trans 1, retra 5, passive disabled, ptop disabled
        key empty, mdkey 1
102: 10.0.102.1,      area index  0, prio 1, cost 1, enabled
    hello 10, dead 40, trans 1, retra 5, passive disabled, ptop disabled
        key empty, mdkey 1

Current OSPF Route Redistribution settings
  Fixed Route Maps:
  Export all with route metric 2, type 1

Current OSPF MD5 key settings:
  1: key encrypted
ACC-1#
```

   Show information about the OSPF-formed adjacencies by running **show ip ospf neighbor** (Example 5-54).

Example 5-54 show ip ospf neighbor command output

```
ACC-1#show ip ospf neighbor
Intf  NeighborID       Prio  State         Address
----  ----------       ----  -----         -------
 101  1.1.1.1             1  Full           10.0.101.2
```

```
 102  1.1.1.2            1  Full          10.0.102.2
ACC-1#
```

Show the OSPF database information by running **show ip ospf database database** (Example 5-55).

Example 5-55 show ip ospf database database command output

```
ACC-1#show ip ospf database

               AS External LSAs

  Link ID         ADV Router      Options Age   Seq#         Checksum
  10.0.10.0       2.2.2.1         0x2     1683  0x80000001   0x97DD
  10.0.10.0       2.2.2.2         0x2     1634  0x80000001   0xE18A
  10.0.30.0       1.1.1.1         0x2     1703  0x80000001   0xE47E
  10.0.30.0       1.1.1.2         0x2     1652  0x80000001   0xEA75
  1.1.1.2         1.1.1.2         0x2     1652  0x80000001   0x8003
  1.1.1.1         1.1.1.1         0x2     1703  0x80000001   0x8403
  2.2.2.2         2.2.2.2         0x2     1634  0x80000001   0x7602
  2.2.2.1         2.2.2.1         0x2     1683  0x80000001   0x364C
  10.0.20.0       2.2.2.1         0x2     1683  0x80000001   0x2942
  10.0.20.0       2.2.2.2         0x2     1634  0x80000001   0x73EE

               Router LSAs (Area 0.0.0.0)

  Link ID         ADV Router      Options Age   Seq#         Checksum
  1.1.1.2         1.1.1.2         0x2     1606  0x80000004   0xD99E
  1.1.1.1         1.1.1.1         0x2     1634  0x80000007   0x3F3E
  2.2.2.2         2.2.2.2         0x2     1602  0x80000004   0x2E31
  2.2.2.1         2.2.2.1         0x2     1620  0x80000005   0x75F5

               Network LSAs (Area 0.0.0.0)

  Link ID         ADV Router      Options Age   Seq#         Checksum
  10.0.102.1      2.2.2.1         0x2     1620  0x80000001   0xC018
  10.0.103.1      1.1.1.1         0x2     1646  0x80000001   0xD901
  10.0.104.2      2.2.2.2         0x2     1602  0x80000002   0xA230
  10.0.100.1      1.1.1.1         0x2     1663  0x80000001   0xD60A
  10.0.101.1      2.2.2.1         0x2     1650  0x80000002   0xBB1E
ACC-1#
```

Show the IP routes learned by OSPF by running **show ip ospf routes** (Example 5-56).

Example 5-56 show ip ospf routes command output

```
ACC-1#show ip ospf routes
Codes: IA - OSPF inter area,
       N1 - OSPF NSSA external type 1, N2 - OSPF NSSA external type 2
       E1 - OSPF external type 1, E2 - OSPF external type 2
       *  - best

    10.0.101.0/30 directly connected
    10.0.102.0/30 directly connected
*E1 1.1.1.1/32 via 10.0.101.2
*E1 1.1.1.2/32 via 10.0.102.2
*E1 2.2.2.2/32 via 10.0.101.2
```

```
*E1 2.2.2.2/32 via 10.0.102.2
 E1 10.0.10.0/24 via 10.0.101.2
 E1 10.0.10.0/24 via 10.0.102.2
 E1 10.0.20.0/24 via 10.0.101.2
 E1 10.0.20.0/24 via 10.0.102.2
*E1 10.0.30.0/24 via 10.0.101.2
*   10.0.100.0/30 via 10.0.101.2
*   10.0.100.0/30 via 10.0.102.2
*   10.0.103.0/30 via 10.0.101.2
*   10.0.104.0/30 via 10.0.102.2
ACC-1#
```

Show the global routing table by running **show ip route** (Example 5-57).

Example 5-57 show ip route command output

```
ACC-1#show ip route
Mgmt routes:
Status code: * - best
    Destination         Mask          Gateway        Type      Tag    Metric   If
  --------------- --------------- --------------- --------- --------- -------- --
  0.0.0.0         0.0.0.0         172.25.1.1      indirect  static             127
* 172.25.0.0      255.255.0.0     172.25.101.122  direct    fixed              127
* 172.25.101.122  255.255.255.255 172.25.101.122  local     addr               127
* 172.25.101.255  255.255.255.255 172.25.101.255  broadcast broadcast          127

Data routes:
Status code: * - best
    Destination         Mask          Gateway        Type      Tag    Metric   If
  --------------- --------------- --------------- --------- --------- -------- --
* 1.1.1.1         255.255.255.255 10.0.101.2      indirect  ospf             3 101
* 1.1.1.2         255.255.255.255 10.0.102.2      indirect  ospf             6 102
* 2.2.2.1         255.255.255.255 2.2.2.1         local     addr               lo1
  2.2.2.1         255.255.255.255 2.2.2.1         broadcast broadcast          lo1
  2.2.2.1         255.255.255.255 2.2.2.1         direct    fixed              lo1
* 2.2.2.2         255.255.255.255 10.0.101.2      indirect  ospf            12 101
* 2.2.2.2         255.255.255.255 10.0.102.2      indirect  ospf            12 102
* 10.0.10.0       255.255.255.0   10.0.10.2       direct    fixed              10
  10.0.10.0       255.255.255.0   10.0.101.2      indirect  ospf            12 101
* 10.0.10.2       255.255.255.255 10.0.10.2       local     addr               10
* 10.0.10.255     255.255.255.255 10.0.10.255     broadcast broadcast          10
  10.0.20.0       255.255.255.0   10.0.101.2      indirect  ospf            12 101
* 10.0.20.0       255.255.255.0   10.0.20.2       direct    fixed              20
* 10.0.20.2       255.255.255.255 10.0.20.2       local     addr               20
* 10.0.20.255     255.255.255.255 10.0.20.255     broadcast broadcast          20
* 10.0.30.0       255.255.255.0   10.0.101.2      indirect  ospf             3 101
* 10.0.100.0      255.255.255.252 10.0.101.2      indirect  ospf             2 101
* 10.0.100.0      255.255.255.252 10.0.102.2      indirect  ospf             2 102
* 10.0.101.0      255.255.255.252 10.0.101.1      direct    fixed              101
* 10.0.101.1      255.255.255.255 10.0.101.1      local     addr               101
* 10.0.101.3      255.255.255.255 10.0.101.3      broadcast broadcast          101
* 10.0.102.0      255.255.255.252 10.0.102.1      direct    fixed              102
* 10.0.102.1      255.255.255.255 10.0.102.1      local     addr               102
* 10.0.102.3      255.255.255.255 10.0.102.3      broadcast broadcast          102
* 10.0.103.0      255.255.255.252 10.0.101.2      indirect  ospf             2 101
* 10.0.104.0      255.255.255.252 10.0.102.2      indirect  ospf             2 102
* 127.0.0.0       255.0.0.0       0.0.0.0         martian   martian
* 224.0.0.0       224.0.0.0       0.0.0.0         martian   martian
* 224.0.0.0       240.0.0.0       0.0.0.0         multicast addr
* 224.0.0.2       255.255.255.255 0.0.0.0         multicast addr
```

```
* 224.0.0.5       255.255.255.255 0.0.0.0         multicast addr
* 224.0.0.6       255.255.255.255 0.0.0.0         multicast addr
* 224.0.0.18      255.255.255.255 0.0.0.0         multicast addr
* 255.255.255.255 255.255.255.255 255.255.255.255 broadcast broadcast
ACC-1#
```

OSPFv3 configuration

OSPF version 3 is based on OSPF version 2, but is modified to support IPv6 addressing. In most other ways, OSPFv3 is similar to OSPFv2: They both have the same packet types and interfaces, and both use the same mechanisms for neighbor discovery, adjacency formation, LSA flooding, aging, and so on. The administrator should be familiar with the OSPFv2 concepts covered in “OSPFv2 configuration” on page 204 before implementing the OSPFv3 differences described in this section.

Although OSPFv2 and OSPFv3 are similar, they represent independent features on the switch. They are configured separately, and both can run in parallel on the switch with no relation to one another, serving different IPv4 (OSPFv2) and IPv6 (OSPFv3) traffic.

OSPFv3 is designed to support IPv6 addresses. IPv6 interfaces must be configured on the switch and assigned to OSPF areas, in much the same way IPv4 interfaces are assigned to areas in OSPFv2. This difference is the primary configuration difference between OSPFv3 and OSPFv2.

For IPv6 interface instructions, see 5.3.2, “Basic IPv6 configuration” on page 192.

The OSPF configuration starts from the assumption that all the configuration tasks presented so far in this chapter have been performed. All the ports are up and running, all the VLANs are correctly defined, all the IPv6 interfaces are configured, the routers are reachable on the Layer 3 links, and the servers can **`ping`** the default gateway.

The goal of the configuration is to enable OSPFv3, to configure the Layer 3 links in the protocol, and to redistribute the connected networks (server subnet) to the neighbors.

The final result is that a route to SRV-1 (FC10::10/64) is installed in the AGG-1 and AGG-2 routing table, a route to SRV-3 (FC30::30/64) is installed in the ACC-1 and ACC-2 routing table, and SRV-1 and SRV-3 are able to ping each other.

To perform a basic OSPFv3 configuration, complete the following steps. For more details and options, see 5.5, “More information” on page 238.

1. Define the Router ID.

   Where OSPFv2 uses a mix of IPv4 interface addresses and Router IDs to identify neighbors, depending on their type, OSPFv3 configuration consistently uses a Router ID to identify all neighbors. Although Router IDs are written in dotted decimal notation, and might even be based on IPv4 addresses from an original OSPFv2 network configuration, it is important to realize that Router IDs are not IP addresses in OSPFv3, and can be assigned independently of the IP address space. However, maintaining Router IDs consistent with any OSPFv2 IPv4 addressing allows for easier implementation of both protocols.

   The same Router IDs are used for both OSPFv2 and OSPFv3, that is, the IP addresses of Loopback 1 interfaces.

2. Enable OSPF.

   Run **ipv6 router ospf** to enter the protocol configuration mode (Example 5-58).

   Example 5-58 OSPF activation

```
ACC-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-1(config)#ipv6 router ospf
ACC-1(config-router-ospf3)#router-id 2.2.2.1
ACC-1(config-router-ospf3)#enable
ACC-1(config-router-ospf3)#^Z
ACC-1#
```

   Important: This operation is performed for all four Layer 3 switches (AGG-1, AGG-2, ACC-1, and ACC-2)

3. Define areas.

   If you are configuring multiple areas in your OSPF domain, one of the areas must be designated as area 0, known as the backbone. The backbone is the central OSPF area and is physically connected to all other areas. The areas inject routing information into the backbone which, in turn, disseminates the information into other areas.

   We use only area 0 (backbone) in our example.

   Important: When OSPFv3 is enabled, the OSPF backbone area (0.0.0.0) is created by default and is always active.

   Up to *three OSPFv3 areas* can be connected to the switch with IBM Networking OS. To configure an area, the OSPF number must be defined and then attached to a network interface on the switch. The full process is explained in the following sections.

   An OSPF area is defined by assigning two pieces of information: an area index and an area ID. The commands to define and enable an OSPF area are:

   - **area <area index> area-id <n.n.n.n>**
   - **area <area index> enable**

   Note: The **area** option is an arbitrary index used only on the switch and does not represent the actual OSPF area number. The actual OSPF area number is defined in the **area-id** parameter of the command.

   The **area <area index>** parameter is an arbitrary index (0 - 2) used only by the switch. This index number does not necessarily represent the OSPF area number, though for configuration simplicity, it should do so where possible.

   For more details about **area-id** formats and options, see 5.5, “More information” on page 238.

4. Attach an area to a network.

   After an OSPF area is defined, it must be associated with a network. To attach the area to a network, you must assign the OSPF area index to an IP interface that participates in the area. Run **ipv6 ospf area <index>** and **ipv6 ospf enable** in interface configuration mode (Example 5-51 on page 207).

   Example 5-59 Attach an area to a network

```
ACC-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-1(config)#interface ip 111
ACC-1(config-ip-if)#ipv6 ospf area 0
ACC-1(config-ip-if)#ipv6 ospf enable
ACC-1(config-ip-if)#^Z
ACC-1#
```

   Important: This operation is performed for all four Layer 3 switches (AGG-1, AGG-2, ACC-1, and ACC-2) for the IP interfaces 110, 111, 112, 113, and 114.

5. Configure route redistribution.

   We use basic redistribution of directly connected (fixed) networks in OSPF. For more details and advanced configuration, see 5.5, “More information” on page 238.

   Run **redistribute {connected|static} export <metric> <AS external metric type>** to redistribute routes (Example 5-60).

   Use equal metric for AGG-1 and ACC-1 and higher metrics for AGG-2 and ACC-2, because we want the traffic from SRV-1 to SRV-3 to go through AGG-1 and ACC-1.

   Example 5-60 Connected routes redistribution

```
AGG-1#conf t
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-1(config)#ipv6 router ospf
AGG-1(config-router-ospf3)#redistribute connected export 2 1
AGG-1(config-router-ospf3)#^Z
AGG-1#

ACC-1#conf t
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-1(config)#ipv6 router ospf
ACC-1(config-router-ospf3)#redistribute connected export 2 1
ACC-1(config-router-ospf3)#^Z
ACC-1#

AGG-2#conf t
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-2(config)#ipv6 router ospf
AGG-2(config-router-ospf3)#redistribute connected export 5 1
AGG-2(config-router-ospf3)#^Z
AGG-2#

ACC-2#conf t
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-2(config)#ipv6 router ospf
ACC-2(config-router-ospf3)#redistribute connected export 10 1
```

```
ACC-2(config-router-ospf3)#^Z
ACC-2#
```

6. Verify OSPF.

   Run the following commands to verify OSPF operation:

   - **show ipv6 ospf**
   - **show ipv6 ospf neighbor**
   - **show ipv6 ospf database**
   - **show ipv6 ospf routes**
   - **show ipv6 route**

   See the following examples for the ACC-1 switch for these commands' output.

   Show the current OSPFv3 configuration settings by running **show ipv6 ospf** (Example 5-61).

Example 5-61 Show ipv6 ospf command output

```
ACC-1#show ipv6 ospf
Current OSPFv3 settings: ON
  Router ID: 2.2.2.1
  lsdb limit none,exit overflow interval 0, ref bw 100000
  spf delay 5, spf holdtime 10, abr type standard
  nssaAsbrDfRtTrans disabled

Current OSPFv3 area settings:
  0: 0.0.0.0,         type transit, transl role candidate, default metric 1
type 1, stb 40, nosumm disabled, enabled

Current OSPFv3 interface settings:
111: fc11:0:0:0:0:0:0:1/64,                  area index  0, instance   0,
enabled
    prio 1, cost 1, hello 10, dead 40, trans delay 1, passive disabled, retra 5
Current IPSec AH protocol: disabled
AH algorithm: none, key , spi 0
Current IPSec ESP protocol: disabled
ESP authentification algorithm none, authentification key
Encryption algorithm null, encryption key , spi 0

112: fc12:0:0:0:0:0:0:1/64,                  area index  0, instance   0,
enabled
    prio 1, cost 1, hello 10, dead 40, trans delay 1, passive disabled, retra 5
Current IPSec AH protocol: disabled
AH algorithm: none, key , spi 0
Current IPSec ESP protocol: disabled
ESP authentification algorithm none, authentification key
Encryption algorithm null, encryption key , spi 0


Current OSPFv3 Route Redistribution settings
  Connected  Export all with route metric 2, type 1, tag unset
ACC-1#
```

Show information about OSPFv3-formed adjacencies by running **show ipv6 ospf neighbor** (Example 5-62).

Example 5-62 show ipv6 ospf neighbor command output

```
ACC-1#show ipv6 ospf neighbor

ID               Pri      State             DeadTime       Address
   1.1.1.1         1        FULL/DR           33
fe80::a17:f4ff:fe32:c46e
   1.1.1.2         1        FULL/DR           38
fe80::fecf:62ff:fe9d:9a6f
ACC-1#
```

Show OSPFv3 database information by running **show ipv6 ospf database** (Example 5-63).

Example 5-63 show ipv6 ospf database command output

```
ACC-1#show ipv6 ospf database

               Router LSAs (Area 0.0.0.0)

  RtrId           Age  Seq#         Checksum   Fragment ID #Links Bits
  1.1.1.1         1200 0x80000084   0xb804     0          3       E
  1.1.1.2         1200 0x80000026   0xcf46     0          3       E
  2.2.2.1         1200 0x8000004a   0xa348     0          2       E
  2.2.2.2         900  0x80000083   0x287f     0          2       E

               Network LSAs (Area 0.0.0.0)

  RtrId           Age  Seq#         Checksum   Link ID     #Rtr
  1.1.1.2         1200 0x8000000a   0x89b      110         2
  2.2.2.1         1500 0x80000012   0xe9ac     111         2
  2.2.2.1         1200 0x80000012   0xeda6     112         2
  2.2.2.2         900  0x80000012   0xd9b8     113         2
  2.2.2.2         900  0x80000012   0xddb2     114         2

               Link LSAs (Area 0.0.0.0)

  RtrId           Age  Seq#         Checksum   Link ID     Intf
  1.1.1.1         1200 0x80000013   0xbd9d     111         111
  2.2.2.1         1500 0x80000013   0x1cc      111         111
  1.1.1.2         900  0x80000013   0xfcfd     112         112
  2.2.2.1         1200 0x80000013   0x2d9d     112         112

               Intra Area Prefix LSAs (Area 0.0.0.0)

  RtrId           Age  Seq#         Checksum   Link ID     Ref-lstype  Ref-LSID
  1.1.1.2         1200 0x80000015   0xf249     0           0x2002      110
  2.2.2.1         900  0x80000089   0x6d44     0           0x2002      111
  2.2.2.1         900  0x80000047   0x6ea      1           0x2002      112
  2.2.2.2         600  0x80000089   0xb1f9     0           0x2002      113
  2.2.2.2         600  0x80000048   0x48a1     1           0x2002      114

               AS-External LSAs
```

```
  RtrId           Age  Seq#         Checksum   Prefix
  1.1.1.1         1200 0x80000012   0xa205     fc11::/64
  1.1.1.1         600  0x80000012   0xb0f3     fc13::/64
  1.1.1.1         600  0x80000011   0xfb8a     fc30::/64
  1.1.1.1         1200 0x8000000a   0xbdfe     fc00::/64
  1.1.1.2         300  0x80000013   0xd6cd     fc12::/64
  1.1.1.2         300  0x80000013   0xe4bc     fc14::/64
  1.1.1.2         300  0x80000012   0x2460     fc30::/64
  1.1.1.2         300  0x8000000b   0xe5d4     fc00::/64
  2.2.2.1         0    0x80000046   0x5c1c     fc10::/64
  2.2.2.1         1500 0x80000012   0x9e08     fc11::/64
  2.2.2.1         1200 0x80000012   0xa004     fc12::/64
  2.2.2.2         300  0x80000049   0x263d     fc10::/64
  2.2.2.2         300  0x80000013   0x98fa     fc13::/64
  2.2.2.2         300  0x80000013   0x9af6     fc14::/64
ACC-1#
```

Show IP routes learned by OSPFv3 by running **show ipv6 ospf routes** (Example 5-64).

Example 5-64 show ipv6 ospf routes command output

```
ACC-1#show ipv6 ospf routes
OSPFV3 Process Routing Table
Dest/Prefix-Length
NextHop/IfIndex                              Cost     Rt.Type   Area
fc00::/64
  fe80::a17:f4ff:fe32:c46e/111                 2        intraArea 0.0.0.0
fc00::/64
  fe80::fecf:62ff:fe9d:9a6f/112                2        intraArea 0.0.0.0
fc10::/64
  fe80::fecf:62ff:fe9d:9a6f/112                12       type1Ext  0.0.0.0
fc10::/64
  fe80::a17:f4ff:fe32:c46e/111                 12       type1Ext  0.0.0.0
fc11::/64
  ::/111                                       1        intraArea 0.0.0.0
fc12::/64
  ::/112                                       1        intraArea 0.0.0.0
fc13::/64
  fe80::a17:f4ff:fe32:c46e/111                 2        intraArea 0.0.0.0
fc14::/64
  fe80::fecf:62ff:fe9d:9a6f/112                2        intraArea 0.0.0.0
fc30::/64
  fe80::a17:f4ff:fe32:c46e/111                 3        type1Ext  0.0.0.0
ACC-1#
```

Show the global routing table by running **show ipv6 route** (Example 5-65).

Example 5-65 show ipv6 route command output

```
ACC-1#show ipv6 route
IPv6 Routing Table  -  11 entries
Codes : C - Connected, S - Static
        O - OSPF
        M - Management Gateway, E - Ext-Management Gateway
O   fc00::/64   [2/110]
        via fe80::a17:f4ff:fe32:c46e, Interface 111
O   fc00::/64   [2/110]
```

```
        via fe80::fecf:62ff:fe9d:9a6f, Interface 112
C   fc10::/64   [1/1]
        via ::, Interface 106
C   fc11::/64   [1/1]
        via ::, Interface 111
C   fc12::/64   [1/1]
        via ::, Interface 112
O   fc13::/64   [2/110]
        via fe80::a17:f4ff:fe32:c46e, Interface 111
O   fc14::/64   [2/110]
        via fe80::fecf:62ff:fe9d:9a6f, Interface 112
O   fc30::/64   [3/110]
        via fe80::a17:f4ff:fe32:c46e, Interface 111
C   fe80::a17:f4ff:fe34:4d69/128   [1/1]
        via ::, Interface 106
C   fe80::a17:f4ff:fe34:4d6e/128   [1/1]
        via ::, Interface 111
C   fe80::a17:f4ff:fe34:4d6f/128   [1/1]
        via ::, Interface 112
ACC-1#
```

OSPFv2 and OSPFv3 verification

Now that the network is routing both IPv4 and IPv6, it is time to test the routes. SRV-1 and SRV-3 should be able to **ping** each other on both IPv4 and IPv6. The details of the SRV-1 and SRV-3 are as follows:

- SRV-1: 10.0.10.10 / FC10::10 (Windows Server 2008 R2)
- SRV-3: 10.0.30.30 / FC30::30 (Red Hat Enterprise Linux)

Windows host verification

Figure 5-11 shows that the Windows host is able to **ping** the Linux host by using both IPv4 and IPv6.

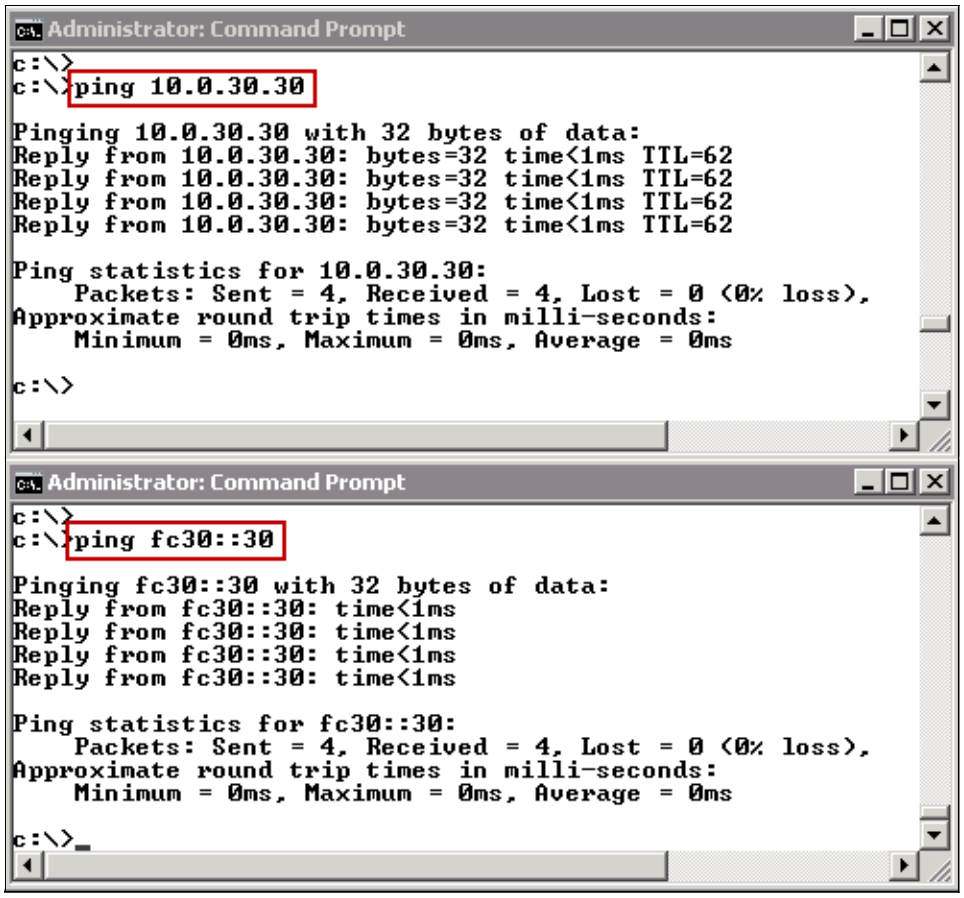

Figure 5-11 Windows host to Linux host verification

Linux host verification

Figure 5-12 shows the output from an ICMP test from a Linux host to a Windows host.

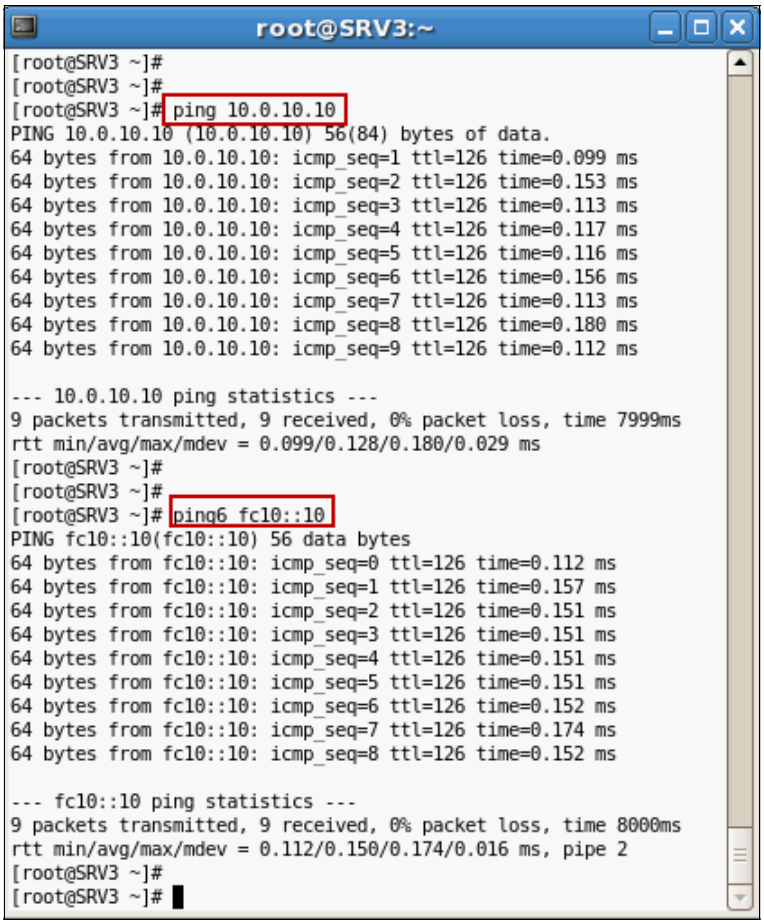
root@SRV3:~
[root@SRV3 ~]#
[root@SRV3 ~]#
[root@SRV3 ~]# ping 10.0.10.10
PING 10.0.10.10 (10.0.10.10) 56(84) bytes of data.
64 bytes from 10.0.10.10: icmp_seq=1 ttl=126 time=0.099 ms
64 bytes from 10.0.10.10: icmp_seq=2 ttl=126 time=0.153 ms
64 bytes from 10.0.10.10: icmp_seq=3 ttl=126 time=0.113 ms
64 bytes from 10.0.10.10: icmp_seq=4 ttl=126 time=0.117 ms
64 bytes from 10.0.10.10: icmp_seq=5 ttl=126 time=0.116 ms
64 bytes from 10.0.10.10: icmp_seq=6 ttl=126 time=0.156 ms
64 bytes from 10.0.10.10: icmp_seq=7 ttl=126 time=0.113 ms
64 bytes from 10.0.10.10: icmp_seq=8 ttl=126 time=0.180 ms
64 bytes from 10.0.10.10: icmp_seq=9 ttl=126 time=0.112 ms
--- 10.0.10.10 ping statistics ---
9 packets transmitted, 9 received, 0% packet loss, time 7999ms
rtt min/avg/max/mdev = 0.099/0.128/0.180/0.029 ms
[root@SRV3 ~]#
[root@SRV3 ~]#
[root@SRV3 ~]# ping6 fc10::10
PING fc10::10(fc10::10) 56 data bytes
64 bytes from fc10::10: icmp_seq=0 ttl=126 time=0.112 ms
64 bytes from fc10::10: icmp_seq=1 ttl=126 time=0.157 ms
64 bytes from fc10::10: icmp_seq=2 ttl=126 time=0.151 ms
64 bytes from fc10::10: icmp_seq=3 ttl=126 time=0.151 ms
64 bytes from fc10::10: icmp_seq=4 ttl=126 time=0.151 ms
64 bytes from fc10::10: icmp_seq=5 ttl=126 time=0.151 ms
64 bytes from fc10::10: icmp_seq=6 ttl=126 time=0.152 ms
64 bytes from fc10::10: icmp_seq=7 ttl=126 time=0.174 ms
64 bytes from fc10::10: icmp_seq=8 ttl=126 time=0.152 ms
--- fc10::10 ping statistics ---
9 packets transmitted, 9 received, 0% packet loss, time 8000ms
rtt min/avg/max/mdev = 0.112/0.150/0.174/0.016 ms, pipe 2
[root@SRV3 ~]#
[root@SRV3 ~]#

Figure 5-12 Linux host to Windows host verification

5.3.3 Border Gateway Protocol

Border Gateway Protocol (BGP) is not in the scope of the reference architecture implementation of this publication. However, a summary of commands used for configuration and verification is presented in this section.

For more information about BGP, see 2.3.5, “Border Gateway Protocol” on page 66.

BGP is an Internet protocol that enables routers on a network to share routing information with each other and advertise information about the segments of the IP address space they can access within their network with routers on external networks. You can use BGP to decide what is the “best” route for a packet to take from your network to a destination on another network, rather than setting a default route from your border routers to your upstream providers. You can configure BGP either within an autonomous system or between different autonomous systems. When run within an autonomous system, it is called internal BGP (iBGP). When run between different autonomous systems, it is called external BGP (eBGP). BGP is defined in RFC 1771 (http://www.ietf.org/rfc/rfc1771.txt).

You can use BGP commands to configure the switch to receive routes and to advertise static routes, fixed routes, and virtual server IP addresses with other internal and external routers. In the current IBM Networking OS implementation, the RackSwitch G8264 switch does not advertise BGP routes that are learned from one iBGP speaker to another iBGP speaker.

Important: BGP is turned off by default.

BGP global configuration

To set the BGP global configuration when using IBM Networking OS, run the following commands:

- Run **`router bgp`** to enter the router BGP configuration mode.
- Run **`neighbor <peer number (1-16)>`** at the BGP level to configure each BGP peer. Each border router, within an autonomous system, exchanges routing information with routers on other external networks.
- Run **`as <0-65535>`** at the BGP level to set the Autonomous System number.
- Run **`local-preference <0-4294967294>`** at the BGP level to set the local preference. The path with the higher value is preferred. When multiple peers advertise the same route, use the route with the shortest AS path as the preferred route if you are using eBGP, or use the local preference if you are using iBGP.
- Run **`maximum-paths <0-32>`** at the BGP level to set maximum paths allowed for an external route. By default, BGP installs only one path to the IP routing table.
- Run **`maximum-paths ibgp <0-32>`** at the BGP level to set the maximum paths allowed for an internal route. By default, BGP installs only one path to the IP routing table.
- Run **`[no] enable`** at the BGP level to globally turn on or off BGP.
- Run **`show ip bgp`** to display the current BGP configuration.

BGP peer configuration

Run the following commands to configure BGP peers, which are border routers that exchange routing information with routers on internal and external networks. The peer option is disabled by default.

- Run **`neighbor <peer number> remote-address <IP address>`** at the BGP level to define the IP address for the specified peer (border router), using dotted decimal notation. The default address is 0.0.0.0.
- Run **`neighbor <peer number> remote-as <1-65535>`** at the BGP level to set the remote autonomous system number for the specified peer.
- Run **`neighbor <peer number> timers hold-time <0, 3-65535>`** at the BGP level to set the time, in seconds, that elapse before the peer session is torn down because the switch does not receive a “keep alive” message from the peer. The default value is 180 seconds.
- Run **`neighbor <peer number> timers keep-alive <0, 1-21845>`** at the BGP level to set the keep-alive time for the specified peer, in seconds. The default value is 60 seconds.
- Run **`neighbor <peer number> advertisement-interval <1-65535>`** at the BGP level to set the time, in seconds, between advertisements. The default value is 60 seconds.
- Run **`neighbor <peer number> retry-interval <1-65535>`** at the BGP level to set the connection retry interval, in seconds. The default value is 120 seconds.

- Run **neighbor <peer number> route-origination-interval <1-65535>** at the BGP level to set the minimum time between route originations, in seconds. The default value is 15 seconds.
- Run **neighbor <peer number> time-to-live <1-255>** at the BGP level to configure the TTL value for a specified peer.

  Time-to-live (TTL) is a value in an IP packet that tells a network router whether the packet has been in the network too long and should be discarded. TTL specifies a certain time span in seconds that, when exhausted, causes the packet to be discarded. The TTL is determined by the number of router hops the packet is allowed before it must be discarded.

  This command specifies the number of router hops that the IP packet can make. This value is used to restrict the number of “hops” the advertisement makes. It is also used to support multi-hops, which allow BGP peers to talk across a routed network. The default number is set at 1.

  Important: The TTL value is significant only to eBGP peers; for iBGP peers, the TTL value in the IP packets is always 255 (regardless of the configured value).

- Run **[no] neighbor <peer number> route-map in <1-32>** at the BGP level to add/remove a route map into the in-route map list.
- Run **[no] neighbor <peer number> route-map out <1-32>** at the BGP level to add/remove a route map into an out-route map list.
- Run **[no] neighbor <peer number> shutdown** at the BGP level to enable or disable peer configuration.
- Run **no neighbor <peer number>** at the BGP level to delete a peer configuration.
- Run **[no] neighbor <peer number> password <1-16 characters>** to configure a BGP peer password.
- Run **[no] neighbor <peer number> passive** at the BGP level to enable or disable BGP passive mode, which prevents the switch from initiating BGP connections with peers. Instead, the switch waits for the peer to send an open message first.
- Run **show ip bgp neighbor [<peer number>]** at the BGP level to display the current BGP peer configuration.

BGP redistribution configuration

Run the following commands to set the BGP redistribution configuration:

- Run **[no] neighbor <peer number> redistribute default-metric <1-4294967294>** at the BGP level to set the default metric of the advertised routes.
- Run **[no] neighbor <peer number> redistribute default-action {import|originate|redistribute}** at the BGP level to set the default route action. Defaults routes can be configured as import, originate, redistribute, or none.
  - Import: Import these routes.
  - Originate: The switch sends a default route to peers if it does not have any default routes in its routing table.

– Redistribute: Default routes are either configured through default gateway or learned through other protocols and redistributed to peers. If the routes are learned from default gateway configuration, you must enable static routes, because the routes from default gateway are static routes. Similarly, if the routes are learned from a certain routing protocol, you must enable that protocol.

Note: No routes are configured in our configuration.

- Run **[no] neighbor <peer number> redistribute rip** at the BGP level to enable or disable advertising RIP routes.
- Run **[no] neighbor <peer number> redistribute ospf** at the BGP level to enable or disable advertising OSPF routes.
- Run **[no] neighbor <peer number> redistribute fixed** at the BGP level to enable or disable advertising fixed routes.
- Run **[no] neighbor <peer number> redistribute static** at the BGP level to enable or disable advertising static routes.
- Run **show ip bgp neighbor <peer number> redistribute** to show the current redistribution configuration.

BGP aggregation configuration

You can use the following commands to configure BGP aggregation to specify the routes/range of IP destinations a peer router accepts from other peers. All matched routes are aggregated to one route, to reduce the size of the routing table. By default, the first aggregation number is enabled and the rest are disabled.

- Run **aggregate-address <1-16> <IP address> <IP netmask>** at the BGP level to define the starting subnet IP address for this aggregation, using dotted decimal notation. The default address is 0.0.0.0.
- Run **[no]aggregate-address <1-16> enable** at the BGP level to enable or disable this BGP aggregation.
- Run **no aggregate-address <1-16>** at the BGP level to delete this BGP aggregation.
- Run **show ip bgp aggregate-address [<1-16>]** to show the current BGP aggregation configuration.

5.4 High availability

This section describes high availability mechanisms in switches running IBM Networking OS. All the features, except Hot Links, were used in the reference architecture, and the implementation steps, along with the commands and their output, are shown in this section.

The topics described in this section are:

- Virtual Router Redundancy Protocol (VRRP)
- Layer 2 Failover
- Trunking
- Hot Links

5.4.1 Virtual Router Redundancy Protocol

The RackSwitch G8264 and RackSwitch G8124 switches support IPv4 high-availability network topologies through an enhanced implementation of VRRP.

VRRP enables redundant router configurations within a LAN, providing alternative router paths for a host to eliminate single points of failure within a network. Each participating VRRP-capable routing device is configured with the same virtual router IPv4 address and ID number. One of the virtual routers is selected as the master, based on a number of priority criteria, and assumes control of the shared virtual router IPv4 address. If the master fails, one of the backup virtual routers takes control of the virtual router IPv4 address and actively processes traffic addressed to it.

For detailed information about VRRP concepts and components, see 2.7.6, “Virtual Router Redundancy Protocol” on page 82.

VRRP is configured on ACC-1 and ACC-2 G8124 access switches only, to provide gateway redundancy for SRV-1.

Complete the following steps to enable, configure, and verify VRRP in IBM Networking OS switches:

1. Enable VRRP.

   Run **`router vrrp`** to enter the VRRP configuration mode and **`enable`** to activate the protocol on both ACC-1 and ACC-2 switches (Example 5-66).

   Example 5-66 Enable VRRP

```
ACC-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-1(config)#router vrrp
ACC-1(config-vrrp)#enable

ACC-2#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-2(config)#router vrrp
ACC-2(config-vrrp)#enable
```

2. Configure the virtual router.

   Complete the following steps to configure the virtual router:

   a. Configure the virtual router ID (Example 5-67).

      Up to 15 virtual router instances can be defined in IBM Networking OS 6.8.

      Run **`virtual-router <1-15> virtual-router-id <1-255>`** in VRRP configuration mode to define the the virtual router ID (VRID). To create a pool of VRRP-enabled routing devices that can provide redundancy to each other, each participating VRRP device must be configured with the same virtual router.

      The VRID for standard virtual routers (where the virtual router IP address is not the same as any virtual server) can be any integer 1 - 255. The default value is 1. All VRID values must be unique within the VLAN to which the virtual router’s IP interface belongs.

      Example 5-67 Virtual router ID configuration

```
ACC-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
```

```
ACC-1(config)#router vrrp
ACC-1(config-vrrp)#virtual-router 1 virtual-router-id 10
ACC-1(config-vrrp)#

ACC-2#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-2(config)#router vrrp
ACC-2(config-vrrp)#virtual-router 1 virtual-router-id 10
ACC-2(config-vrrp)#
```

b. Configure the Virtual router IP address (Example 5-68).

Run **[no] virtual-router <1-15> address <IP address>** to define an IP address for this virtual router, using dotted decimal notation. This command is used in conjunction with the VRID to configure the same virtual router on each participating VRRP device. The default address is 0.0.0.0.

Example 5-68 Virtual router IP address configuration

```
ACC-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-1(config)#router vrrp
ACC-1(config-vrrp)#virtual-router 1 address 10.0.10.1
ACC-1(config-vrrp)#

ACC-2#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-2(config)#router vrrp
ACC-2(config-vrrp)#virtual-router 1 address 10.0.10.1
ACC-2(config-vrrp)#
```

c. Select a switch IP interface (Example 5-69).

Run **virtual-router <1-15> interface <interface number>** to select a switch IP interface. If the IP interface has the same IP address as the **addr** option, this switch is considered the “owner” of the defined virtual router. An owner has a special priority of 255 (highest) and always assumes the role of master router, even if it must pre-empt another virtual router that assumed master routing authority. This preemption occurs even if the **preem** option is disabled. The default value is 1.

Example 5-69 IP interface selection

```
ACC-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-1(config)#router vrrp
ACC-1(config-vrrp)#virtual-router 1 interface 10
ACC-1(config-vrrp)#

ACC-2#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-2(config)#router vrrp
ACC-2(config-vrrp)#virtual-router 1 interface 10
ACC-2(config-vrrp)#
```

d. Define the election priority (Example 5-70).

Run **virtual-router <1-15> priority <1-254>** to define the election priority bias for this virtual server. The priority value can be any integer 1 - 254. The default value is 100.

During the master router election process, the routing device with the highest virtual router priority number wins. If there is a tie, the device with the highest IP interface address wins. If this virtual router's IP address is the same as the one used by the IP interface, the priority for this virtual router is automatically set to 255 (highest).

When priority tracking is used, this base priority value can be modified according to a number of performance and operational criteria.

In our reference architecture, we used a different Virtual router IP address from the IP interface addresses and assigned a higher priority to ACC-1 switch, in order for it to become an elected master. The ACC-2 switch is left as its default priority (100) (Example 5-70).

Example 5-70 Virtual router election priority configuration

```
ACC-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-1(config)#router vrrp
ACC-1(config-vrrp)#virtual-router 1 priority 105
ACC-1(config-vrrp)#
```

e. Configure preemption (Example 5-71).

Run **[no] virtual-router <1-15> preemption** to enable or disable master preemption. When enabled, if this virtual router is in backup mode but has a higher priority than the current master, this virtual router pre-empts the lower priority master and assumes control. Even when preemption is disabled, this virtual router always pre-empts any other master if this switch is the owner (the IP interface address and virtual router address are the same). By default, this option is enabled.

Example 5-71 Configure preemption

```
ACC-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-1(config)#router vrrp
ACC-1(config-vrrp)#virtual-router 1 preemption
ACC-1(config-vrrp)#

ACC-2#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-2(config)#router vrrp
ACC-2(config-vrrp)#virtual-router 1 preemption
ACC-2(config-vrrp)#
```

f. Configure the timers (Example 5-72 on page 226).

Run **virtual-router <1-15> timers advertise <1-255>** to define the time interval between VRRP master advertisements. This interval can be any integer 1 - 255 seconds. The default value is 1.

Run **virtual-router <1-15> timers preempt-delay-time <0-255>** to configure the pre-empt delay interval. This timer is configured on the VRRP Owner and prevents the switch from moving back to the Master state until the preemption delay interval expires. Ensure that the interval is long enough for OSPF or other routing protocols to converge.

Run **[no] virtual-router <1-128> fast-advertise** to enable or disable fast advertisements. When enabled, the VRRP master advertisements interval is calculated in units of centiseconds, instead of seconds. For example, if advertisement is set to 1 and fast advertisement is enabled, master advertisements are sent every .01 second. When you disable fast advertisement, the advertisement interval is set to the default value of 1 second. To support Fast Advertisements, set the interval to 20 - 100 centiseconds.

Important: Fast advertisements were not used in the reference architecture implementation.

Example 5-72 Virtual router timers configuration

```
ACC-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-1(config)#router vrrp
ACC-1(config-vrrp)#virtual-router 1 timers advertise 2
ACC-1(config-vrrp)#virtual-router 1 timers preempt-delay-time 5
ACC-1(config-vrrp)#

ACC-2#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-2(config)#router vrrp
ACC-2(config-vrrp)#virtual-router 1 timers advertise 2
ACC-2(config-vrrp)#virtual-router 1 timers preempt-delay-time 5
ACC-2(config-vrrp)#
```

g. Enable the configured virtual router.

Run **[no] virtual-router <1-15> enable** to enable or disable this virtual router.

Run **no virtual-router <1-15>** to delete this virtual router from the switch configuration.

Important: We enabled the ACC-2 (backup) router first to show the preemption operation. As shown in Example 5-73, the log messages show the role status change. Observe the time stamps of the messages. You can see that ACC-1 assumed the Master role and ACC-2 returned to Backup state after the virtual router is enabled on ACC-1 with the preemption option enabled.

Example 5-73 Virtual router activation with preemption enabled

```
ACC-2#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-2(config)#router vrrp
ACC-2(config-vrrp)#virtual-router 1 enable
ACC-2(config-vrrp)#

Aug 15 23:44:56 ACC-2 NOTICE  vrrp: virtual router       10.0.10.1 is now BACKUP
Aug 15 23:45:03 ACC-2 NOTICE  vrrp: virtual router       10.0.10.1 is now MASTER.

Aug 15 23:45:22 ACC-2 NOTICE  vrrp: virtual router       10.0.10.1 is now BACKUP

ACC-2(config-vrrp)#^Z
ACC-2#
ACC-2#
```

```
ACC-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-1(config)#router vrrp
ACC-1(config-vrrp)#virtual-router 1 enable
ACC-1(config-vrrp)#

Aug 15 23:45:18 ACC-1 NOTICE  vrrp: virtual router       10.0.10.1 is now BACKUP
Aug 15 23:45:23 ACC-1 NOTICE  vrrp: virtual router       10.0.10.1 is now MASTER.

ACC-1(config-vrrp)#^Z

ACC-1#
```

3. Configure tracking.

   The commands in this step are used for modifying the priority system used when electing the master router from a pool of virtual routers. Various tracking criteria can be used to bias the election results. Each time one of the tracking criteria is met, the priority level for the virtual router is increased by an amount defined through the VRRP Tracking commands. The criteria are tracked dynamically, continuously updating virtual router priority levels when enabled. If the virtual router preemption option is enabled, this virtual router can assume master routing authority when its priority level rises above the priority of the current master.

   Some tracking criteria apply to standard virtual routers, otherwise called “virtual interface routers.” A virtual server router is defined as any virtual router whose IP address is the same as any configured virtual server IP address.

   – Run **[no] virtual-router <1-15> track virtual-routers** to allow the priority for this virtual router to be increased for each virtual router in Master mode on this switch. This situation is useful for making sure that traffic for any particular client/server pairing is handled by the same switch, increasing routing and load balancing efficiency. This command is disabled by default.

   – Run **[no] virtual-router <1-15> track interfaces** to allow the priority for this virtual router to be increased for each other IP interfaces active on this switch. An IP interface is considered active when there is at least one active port on the same VLAN. This situation helps elect the virtual routers with the most available routes as the master. This command is disabled by default.

   – Run **[no] virtual-router <1-15> track ports** to allow the priority for this virtual router to be increased for each active port on the same VLAN. A port is considered “active” if it has a link and is forwarding traffic. This situation helps elect the virtual routers with the most available ports as the master. This command is disabled by default.

   Run the following commands to set weights for the various criteria used to modify priority levels during the master router election process. Each time one of the tracking criteria is met, the priority level for the virtual router is increased by a defined amount. These priority tracking options define increment values only. These options do not affect the VRRP master router election process until options under the VRRP Virtual Router Priority Tracking Commands are enabled.

   – Run **tracking-priority-increment virtual-routers <0-254>** to define the priority increment value (0 - 254) for virtual routers in Master mode detected on this switch. The default value is 2.

   – Run **tracking-priority-increment interfaces <0-254>** to define the priority increment value for active IP interfaces detected on this switch. The default value is 2.

– Run **tracking-priority-increment ports <0-254>** to define the priority increment value for active ports on the virtual router's VLAN. The default value is 2.

Tracking configuration is shown in Example 5-74.

Example 5-74 Configure tracking

```
ACC-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-1(config)#router vrrp
ACC-1(config-vrrp)#virtual-router 1 track ports
ACC-1(config-vrrp)#tracking-priority-increment ports 50
ACC-1(config-vrrp)#

ACC-2#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-2(config)#router vrrp
ACC-2(config-vrrp)#virtual-router 1 track ports
ACC-2(config-vrrp)#tracking-priority-increment ports 50
ACC-2(config-vrrp)#
```

Important: For other VRRP configuration commands and details, see 5.5, "More information" on page 238.

4. Verify the VRRP operation

Run the commands in this step to verify the VRRP operation on the switch.

Run **show ip vrrp** to show the current VRRP parameters (Example 5-75).

Example 5-75 Verify the VRRP current parameters

```
ACC-1#show ip vrrp
Current VRRP settings: ON
Current VRRP hold off time: 0

Current VRRP Tracking settings:
   vrs 2, ifs 2, ports 50
Current VRRP Virtual Router Group:
    vrid 1, if 1, prio 100, adver 1, disabled
    preem enabled, fast-advertisement disabled
    track nothing
Current VRRP virtual router settings:
  1: vrid 10, 10.0.10.1, if 10, prio 105, adver 2, enabled
    preem enabled, predelay 5, fast-advertisement disabled
    track ports
ACC-1#


ACC-2#show ip vrrp
Current VRRP settings: ON
Current VRRP hold off time: 0

Current VRRP Tracking settings:
   vrs 2, ifs 2, ports 50
Current VRRP Virtual Router Group:
    vrid 1, if 1, prio 100, adver 1, disabled
    preem enabled, fast-advertisement disabled
```

```
    track nothing
Current VRRP virtual router settings:
  1: vrid 10, 10.0.10.1, if 10, prio 100, adver 2, enabled
    preem enabled, predelay 5, fast-advertisement disabled
    track ports
ACC-2#
```

Run **show ip vrrp virtual-router <1-15>** to show the current VRRP parameters of the selected virtual router (Example 5-76).

Example 5-76 Verify the selected virtual router current parameters

```
ACC-1#show ip vrrp virtual-router 1
Current VRRP virtual router 1:
 vrid 10, 10.0.10.1, if 10, prio 105, adver 2, enabled
    preem enabled, predelay 5, fast-advertisement disabled
    track ports
ACC-1#


ACC-2#show ip vrrp virtual-router 1
Current VRRP virtual router 1:
 vrid 10, 10.0.10.1, if 10, prio 100, adver 2, enabled
    preem enabled, predelay 5, fast-advertisement disabled
    track ports
ACC-2#
```

Run **show ip vrrp counters** to show the VRRP statistics (Example 5-77).

Example 5-77 Show VRRP statistics

```
ACC-1#show ip vrrp counters
VRRP statistics:
vrrpInAdvers:         40752   vrrpBadAdvers:                74
vrrpOutAdvers:        46019   vrrpOutGratuitousARPs:         4
vrrpBadVersion:           0   vrrpBadVrid:                  59
vrrpBadAddress:           0   vrrpBadData:                   0
vrrpBadPassword:          0   vrrpBadInterval:              15
ACC-1#
```

Run **show ip vrrp track-priority-increment** to show the VRRP tracking priority increments configuration (Example 5-78).

Example 5-78 Show the VRRP tracking priority increments

```
ACC-1#show ip vrrp track-priority-increment
Current VRRP Tracking settings:
   vrs 2, ifs 2, ports 50
ACC-1#
```

Run **show ip vrrp virtual-router <1-15> track** to show the selected virtual router tracking configuration (Example 5-79).

Example 5-79 Show the virtual router tracking configuration

```
ACC-1#show ip vrrp virtual-router 1 track
Current VRRP virtual router 1 tracking:
```

```
       track ports
ACC-1#
```

5.4.2 Layer 2 Failover

The primary application for Layer 2 Failover is to support Network Adapter Teaming. With Network Adapter Teaming, all the NICs on each server share an IP address, and are configured into a team. One NIC is the primary link, and the other is a standby link. For more details, see the documentation for your Ethernet adapter.

Link limits: Only two links per server can be used for Layer 2 Failover (one primary and one backup). Network Adapter Teaming allows only one backup NIC for each server blade.

Layer 2 Failover can be enabled on any trunk group in the switch, including LACP trunks. Trunks can be added to failover trigger groups. Then, if some specified number of monitor links fail, the switch disables all the control ports in the switch. When the control ports are disabled, it causes the NIC team on the affected servers to fail over from the primary to the backup NIC. This process is called a *failover event*.

When the appropriate number of links in a monitor group return to service, the switch enables the control ports. This action causes the NIC team on the affected servers to fail back to the primary switch (unless Auto-Fallback is disabled on the NIC team). The backup switch processes traffic until the primary switch's control links come up, which can take up to 5 seconds.

For more details about the Layer 2 Failover feature, see Chapter 2, "IBM System Networking Switch 10Gb Ethernet switch features" on page 51.

For our Layer 2 Failover implementation, we used portchannel 1 and portchannel 2 on the ACC-1 and ACC-2 access switches to control the SRV-1 port on the switch (port 7). For a reminder of the topology, see the network diagrams in the Chapter 3, "Reference architectures" on page 107.

Basically, when uplink trunks of ACC-1 or ACC-2 go down, the access switches disable the server port in order for the NIC teaming mechanism on the server to use the other network interface.

Not all the options available in IBM Networking OS were used for our Layer 2 Failover configuration. For detailed information about failover configuration commands, see 5.5, "More information" on page 238.

Complete the following steps to enable, configure, and verify Layer 2 Failover:

1. Enable Layer 2 Failover.

   Run **`failover enable`** to enable Layer 2 failover globally (Example 5-80).

   Example 5-80 Enable Layer 2 Failover on the switch

```
ACC-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-1(config)#failover enable
ACC-1(config)#^Z
ACC-1#
```

2. Configure the Failover Manual Monitor Port.

   Run **[no] failover trigger <1-8> mmon monitor portchannel <trunk number>** command to add or remove the selected trunk group to the Manual Monitor Port configuration (Example 5-81). These ports are the whose states the switch monitors to control the server port link.

Example 5-81 Failover Manual Monitor Port configuration

```
ACC-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-1(config)#failover trigger 1 mmon monitor portchannel 1
ACC-1(config)#failover trigger 1 mmon monitor portchannel 2
ACC-1(config)#^Z
ACC-1#


ACC-2#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-2(config)#failover trigger 1 mmon monitor portchannel 1
ACC-2(config)#failover trigger 1 mmon monitor portchannel 2
ACC-2(config)#^Z
ACC-2#
```

Portchannels: Portchannel 1 and portchannel 2 are the uplink trunks to AGG-1 and AGG-2 of the ACC-1 and ACC-2 switches

3. Configure Failover Manual Monitor Control.

   Run **[no] failover trigger <1-8> mmon control member <port alias or number>** to add or remove the selected port to the Manual Monitor Control configuration (Example 5-82).

   This port is the port the switch controls (the server port).

Example 5-82 Failover Manual Monitor Control configuration

```
ACC-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-1(config)#failover trigger 1 mmon control member 7
ACC-1(config)#^Z
ACC-1#


ACC-2#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-2(config)#failover trigger 1 mmon control member 7
ACC-2(config)#^Z
ACC-2#
```

4. Configure the Failover Trigger limit.

   Run **failover trigger <1-8> limit <0-1024>** to configure the minimum number of operational links allowed within each trigger before the trigger initiates a failover event (Example 5-83). If you enter a value of zero (0), the switch triggers a failover event only when no links in the trigger are operational.

   Example 5-83 Failover Trigger limit configuration

```
ACC-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-1(config)#failover trigger 1 limit 2
ACC-1(config)#^Z
ACC-1#


ACC-2#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-2(config)#failover trigger 1 limit 2
ACC-2(config)#^Z
ACC-2#
```

5. Enable the Failover Trigger.

   Run **[no] failover trigger <1-8> enable** to enable or disable the selected Failover Trigger (Example 5-84).

   Example 5-84 Enable the Failover Trigger

```
ACC-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-1(config)#failover trigger 1 enable

Aug 11 12:44:29 ACC-1 NOTICE  failover: Trigger 1 is up, control ports are auto
controlled.

ACC-1(config)#^Z
ACC-1#


ACC-2#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-2(config)#failover trigger 1 enable

Aug 11 12:44:40 ACC-2 NOTICE  failover: Trigger 1 is up, control ports are auto
controlled.

ACC-2(config)#^Z
ACC-2#

failover trigger 1 enable
```

6. Verify the failover configuration.

   Run **show failover trigger <1-8>** to show the current Failover Trigger settings (Example 5-85).

   Example 5-85 Display current fAilover Trigger settings

```
ACC-1#show failover trigger 1
Current Trigger 1 setting: enabled
limit 2
Manual Monitor settings:
        trunks 1 2
Manual Control settings:
        ports 7
ACC-1#


ACC-2#show failover trigger 1
Current Trigger 1 setting: enabled
limit 2
Manual Monitor settings:
        trunks 1 2
Manual Control settings:
        ports 7
ACC-2#
```

7. Test the Layer 2 Failover operation

   In our reference architecture, ports 1 - 4 on the ACC-1 and ACC-2 access switches are the uplink ports to AGG-1 and AGG-2 switches.

   We shut down two ports out of four on an access switch and verify that the controlled port (port 7 connected to SRV-1) is automatically disabled.

   Important: This example uses collected command output from the ACC-1 switch only. The same commands can be used on ACC-2 as well.

   a. Shut down two uplink ports.

      Run **shutdown** to manually disable the ports and see the messages related to the failover operation (Example 5-86).

      Example 5-86 Manually disable monitored ports

```
ACC-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-1(config)#interface port 1,2
ACC-1(config-if)#shutdown
ACC-1(config-if)#


ACC-1#
Aug 11 12:59:21 ACC-1 NOTICE  link: link down on port 1

Aug 11 12:59:21 ACC-1 NOTICE  link: link down on port 2

Aug 11 12:59:21 ACC-1 WARNING failover: Trigger 1 is down, control ports are
auto disabled.
```

```
Aug 11 12:59:21 ACC-1 NOTICE  link: link down on port 7


ACC-1(config-if)#^Z
ACC-1#
```

b. Verify the Failover Trigger status.

Run **show failover trigger <1-8> information** to verify the failover trigger status (Example 5-87).

Example 5-87 Verify the Failover Trigger status

```
ACC-1#show failover trigger 1 information

Trigger 1 Manual Monitor: Enabled
Trigger 1 limit: 2
Monitor State: Down
Member      Status
---------   -----------
PortChannel 1
 1          Failed
 2          Failed
PortChannel 2
 3          Operational
 4          Operational
Control State: Auto Disabled
Member      Status
---------   -----------
 7          Failed
ACC-1#
```

c. Verify the interface status.

Run **show interface status** to verify that the control port (port 7) is also disabled along with the manually disabled uplink ports (1,2) (Example 5-88).

Example 5-88 Verify the interface status

```
ACC-1#show interface status
------------------------------------------------------------------
Alias   Port   Speed    Duplex     Flow Ctrl      Link
------- ----   -----   --------  --TX-----RX--   ------
1        1      10000     full      no     no    disabled
2        2      10000     full      no     no    disabled
3        3      10000     full      no     no       up
4        4      10000     full      no     no       up
5        5      10000     full      no     no       up
6        6      10000     full      no     no       up
7        7      10000     full      no     no    disabled
8        8      10000     full      no     no       up
9        9      1G/10G    full      no     no      down
10       10     1G/10G    full      no     no      down
11       11     1G/10G    full      no     no      down
12       12     1G/10G    full      no     no      down
13       13     10000     full      no     no       up
14       14     1G/10G    full      no     no      down
```

```
15       15     1G/10G    full      no     no      down
16       16     10000     full      no     no       up
17       17     1G/10G    full      no     no      down
18       18     1G/10G    full      no     no      down
19       19     1G/10G    full      no     no      down
20       20     1G/10G    full      no     no      down
21       21     1G/10G    full      no     no      down
22       22     1G/10G    full      no     no      down
23       23     1G/10G    full      no     no      down
24       24     10000     full      no     no       up
MGTA     25      1000     full     yes    yes       up
MGTB     26       any      any     yes    yes       up
ACC-1#
```

d. Enable the two uplink ports.

Run **no shutdown** to manually enable the ports and see the messages related to the failover operation (Example 5-89).

Example 5-89 Manually enable the ports

```
ACC-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-1(config)#
ACC-1(config)#interface port 1,2
ACC-1(config-if)#no shutdown
ACC-1(config-if)#


Aug 11 13:00:57 ACC-1 NOTICE  link: link up on port 1

Aug 11 13:00:58 ACC-1 NOTICE  link: link up on port 2

Aug 11 13:00:58 ACC-1 NOTICE  failover: Trigger 1 is up, control ports are
auto controlled.

Aug 11 13:01:03 ACC-1 NOTICE  link: link up on port 7


ACC-1(config-if)#^Z
ACC-1#
```

e. Verify the Failover Trigger status.

Run **show failover trigger <1-8> information** to verify the Failover Trigger status (Example 5-90).

Example 5-90 Verify the Failover Trigger status

```
ACC-1#show failover trigger 1 information

Trigger 1 Manual Monitor: Enabled
Trigger 1 limit: 2
Monitor State: Up
Member      Status
---------   -----------
PortChannel 1
 1          Operational
```

```
 2          Operational
PortChannel 2
 3          Operational
 4          Operational
Control State: Auto Controlled
Member      Status
---------   -----------
 7          Operational
ACC-1#
```

f. Verify the interface status.

Run **`show interface status`** to verify that the control port (port 7) is also disabled along with the manually disabled uplink ports (1,2) (Example 5-91).

Example 5-91 Verify the interface status

```
ACC-1#show interface status
------------------------------------------------------------------------
Alias   Port   Speed    Duplex     Flow Ctrl      Link
------- ----   -----   --------  --TX-----RX--   ------
1        1      10000     full      no     no       up
2        2      10000     full      no     no       up
3        3      10000     full      no     no       up
4        4      10000     full      no     no       up
5        5      10000     full      no     no       up
6        6      10000     full      no     no       up
7        7      10000     full      no     no       up
8        8      10000     full      no     no       up
9        9      1G/10G    full      no     no      down
10       10     1G/10G    full      no     no      down
11       11     1G/10G    full      no     no      down
12       12     1G/10G    full      no     no      down
13       13     10000     full      no     no       up
14       14     1G/10G    full      no     no      down
15       15     1G/10G    full      no     no      down
16       16     10000     full      no     no       up
17       17     1G/10G    full      no     no      down
18       18     1G/10G    full      no     no      down
19       19     1G/10G    full      no     no      down
20       20     1G/10G    full      no     no      down
21       21     1G/10G    full      no     no      down
22       22     1G/10G    full      no     no      down
23       23     1G/10G    full      no     no      down
24       24     10000     full      no     no       up
MGTA     25      1000     full     yes    yes       up
MGTB     26       any      any     yes    yes       up
ACC-1#
```

5.4.3 Trunking

Multiple switch ports can be combined together to form robust, high-bandwidth trunks to other devices. Since trunks are composed of multiple physical links, the trunk group is inherently fault tolerant. If connection between the switches is available, the trunk remains active.

For detailed information about trunking, see 5.2.2, “Ports and trunking” on page 170.

5.4.4 Hot Links

Important: The Hot Links function was not used in the reference architecture implementation. The following section is just a short outline of the feature. For more information, see 5.5, “More information” on page 238.

For network topologies that require STP to be turned off, the Hot Links function provides basic link redundancy with fast recovery.

Hot Links consists of up to 25 triggers. A trigger consists of a pair of Layer 2 interfaces, each containing an individual port, trunk, or LACP adminkey. One interface is the Master, and the other is a Backup. While the Master interface is set to the active state and forwards traffic, the Backup interface is set to the standby state and blocks traffic until the Master interface fails. If the Master interface fails, the Backup interface is set to active and forwards traffic. After the Master interface is restored, it moves to the standby state and blocks traffic until the Backup interface fails.

You may select a physical port, static trunk, or an LACP adminkey as a Hot Link interface.

Configuration guidelines

The following configuration guidelines apply to Hot links:

- Ports that are configured as Hot Link interfaces must have STP disabled.
- When Hot Links is turned on, MSTP, RSTP, and PVRST must be turned off.
- When Hot Links is turned on, UplinkFast must be disabled.
- A port that is a member of the Master interface cannot be a member of the Backup interface.
- A port that is a member of one Hot Links trigger cannot be a member of another Hot Links trigger.
- An individual port that is configured as a Hot Link interface cannot be a member of a trunk.

Configuring Hot Links

Run the following commands to configure Hot Links.

- **`switch(config)# hotlinks trigger 1 enable`**: Enables Hot Links trigger 1)
- **`switch(config)# hotlinks trigger 1 master port 1`**: Adds a port to the Master interface.
- **`switch(config)# hotlinks trigger 1 backup port 2`**: Adds a port to the Backup interface.
- **`switch(config)# hotlinks enable`**: Turns on Hot Links.

5.5 More information

For more information about the topics described in this chapter, see the following documentation

- Configuration guides:
  - *IBM RackSwitch G8264 Application Guide (6.8)*:

    http://www-01.ibm.com/support/docview.wss?uid=isg3T7000464
  - *IBM RackSwitch G8124/G8124-E Application Guide (6.8)*:

    http://www-01.ibm.com/support/docview.wss?uid=isg3T7000465
- Command reference guides:
  - *IBM RackSwitch G8264 Menu-based CLI Reference (6.8)*:

    http://www-01.ibm.com/support/docview.wss?uid=isg3T7000481
  - *IBM RackSwitch G8124/G8124-E ISCLI Command Reference (6.6)*:

    https://www-304.ibm.com/support/docview.wss?uid=isg3T7000350

Refer to the IBM Support website for a complete list of the documentation at:

http://www-947.ibm.com/support/entry/portal/Documentation?lnk=mhsd

6

IBM Virtual Fabric 10Gb Switch Module implementation

This chapter provides information and instructions for implementing the IBM Virtual Fabric 10Gb Switch Module for IBM BladeCenter. Using the reference architecture described in Chapter 3, “Reference architectures” on page 107, this chapter presents a step by step guide for implementing and configuring the most important functions available in IBM Networking OS. It is not meant to cover all the features in the operating system. Instead, it strictly refers to the proposed test architecture.

The network topology used for Virtual Fabric 10Gb Switch Module (VFSM) implementation is built by using two switch modules, installed in an IBM BladeCenter H chassis and a blade server that runs the Red Hat Enterprise Linux operating system.

This network configuration is not meant to be a stand-alone network configuration, but integrated with the IBM RackSwitch implementation described in Chapter 5, “IBM System Networking RackSwitch implementation” on page 155.

The goal of this chapter is that the final result is a fully functional, integrated network that is able to prove the operation of the implemented features.

© Copyright IBM Corp. 2012. All rights reserved.

6.1 Purpose of this implementation

This implementation shows how a mixed environment of both stand-alone and embedded switches provide end-to-end communication in a data center, for servers that run different operating systems and IP protocol versions (IPv4 and IPv6).

At the conclusion of this implementation, the Linux host installed in the BladeCenter H chassis is able to communicate with the Windows host connected to the G8124 switches of a Top-of-Rack implementation.

If you have the same (or equivalent) equipment that is used in the reference architecture, you should be able to replicate the described configuration by following the steps presented in this chapter, and arrive at the same result.

For Virtual Fabric 10Gb Switch Module operating in stand-alone mode, the topics described in Chapter 5, “IBM System Networking RackSwitch implementation” on page 155 apply as well.

To have a feature-rich implementation and configuration diversity, the VFSMs use a stacking topology and integrate with the stand-alone equipment by using common features, such as VLANs, trunking, Spanning Tree Protocol (STP), and basic IP routing.

Although stacking VFSMs adds many restrictions in terms of the remaining available software features, it adds value to the reference architecture by presenting configuration aspects that are not covered in the RackSwitch implementation.

When implementing stacking, many advanced IP features, such as OSPF, BGP, VRRP, IPv6, and others, become unavailable. For stacking limitations, see 6.2, “Stacking” on page 240.

For that reason, Virtual Fabric 10Gb Switch Modules are implemented as access switches and their configuration focus mainly on stacking and Layer 2 features, with no Layer 3 functions (performed at the aggregation layer).

The topics described in this chapter are:

- Stacking implementation: Overview, requirements, limitations, configuration, operation, and redundancy
- Layer 1 implementation: Ports connection, configuration, and verification
- Layer 2 implementation: VLANs, tagging, trunking, STP, and QoS
- High availability: Stacking, Layer 2 Failover, trunking, Hot Links, and VRRP at the aggregation layer

6.2 Stacking

If you decide to use stacking with Virtual Fabric 10Gb Switch Modules, then you must first perform stack initialization and configuration.

Note: Do not configure anything on the switch before the stack is set up. Enabling and configuring stacking forces the modules to reboot with the factory default configuration. Ports are renumbered and the stand-alone configuration is no longer valid. Make sure you back up your stand-alone configuration before proceeding to stack initialization.

6.2.1 Stacking overview

A stack is a group of up to eight Virtual Fabric 10Gb Switch Module switches with IBM Networking OS that work together as a unified system. A stack has the following properties, regardless of the number of switches included:

- The network views the stack as a single entity.
- The stack can be accessed and managed as a whole using standard switch IP interfaces configured with IPv4 addresses.
- After the stacking links are established, the number of ports available in a stack equals the total number of remaining ports of all the switches that are part of the stack.
- The number of available IP interfaces, VLANs, trunks, trunk links, and other switch attributes are not aggregated among the switches in a stack. The totals for the stack as a whole are the same as for any single switch configured in stand-alone mode.

6.2.2 Stacking requirements

Before IBM Networking OS switches can form a stack, they must meet the following requirements:

- All switches must be the same model (Virtual Fabric 10Gb Switch Module).
- Each switch must have IBM Networking OS V6.5 or later installed. The same release version is not required, as the Master switch pushes a firmware image to each switch that is part of the stack.
- The preferred stacking topology is a bidirectional ring. To achieve this topology, reserve two external 10 Gb Ethernet ports on each switch for stacking. By default, the first two 10 Gb Ethernet ports are used.
- The cables used for connecting the switches in a stack carry low-level, inter-switch communications and cross-stack data traffic critical to shared switching functions. Always maintain the stability of stack links to avoid internal stack reconfiguration.

6.2.3 Stacking limitations

A VFSM with IBM Networking OS V6.5 and above can operate in one of two modes:

- Default mode, which is the regular stand-alone (or non-stacked) mode.
- Stacking mode, in which multiple physical switches aggregate functions as a single switching device.

When in stacking mode, the following stand-alone features are not supported:

- Active Multi-Path Protocol (AMP)
- BCM rate control
- Border Gateway Protocol (BGP)
- Converge Enhanced Ethernet (CEE)
- Fibre Channel over Ethernet (FCoE)
- IGMP Relay and IGMPv3
- IPv6
- Link Layer Detection Protocol (LLDP)
- Loopback Interfaces
- MAC address notification
- MSTP
- OSPF and OSPFv3

- Port flood blocking
- Protocol-based VLANs
- RIP
- Router IDs
- Route maps
- sFlow port monitoring
- Static MAC address addition
- Static multicast
- Uni-Directional Link Detection (UDLD)
- Virtual NICs
- Virtual Router Redundancy Protocol (VRRP)

Important: In stacking mode, switch menus and commands for unsupported features might be unavailable, or might have no effect on switch operation.

6.2.4 Stack membership

A stack contains up to eight switches, interconnected by a stack trunk in a local ring topology. With this topology, only a single stack link failure is allowed. An operational stack must contain one Master and one or more Members, as follows:

- Master: One switch controls the operation of the stack and is called the Master. The Master provides a single point to manage the stack. A stack must have only one Master. The firmware image, configuration information, and runtime data are maintained by the Master and pushed to each switch in the stack as necessary.
- Member: Member switches provide additional port capacity to the stack. Members receive configuration changes, runtime information, and software updates from the Master.
- Backup: One member switch can be designated as a Backup to the Master. The Backup takes over control of the stack if the Master fails. Configuration information and runtime data are synchronized with the Master.

The Master switch

An operational stack can have only one active Master at any time. In a normal stack configuration, one switch is configured as a Master and all the other switches are configured as Members.

When adding new switches to an existing stack, the administrator should configure each new switch for its intended role as a Master (only when replacing a previous Master) or as a Member. All stack configuration procedures in this chapter show correct role specification.

However, although uncommon, there are scenarios in which a stack may temporarily have more than one Master switch. Should this situation occur, one Master switch is automatically chosen as the active Master for the entire stack. The selection process is designed to promote stable, predictable stack operation and minimize stack reboots and other disruptions.

Splitting and merging one stack

Important: If stack links or Member switches fail, any Members that cannot access either the Master or Backup are considered isolated and do not process network traffic.

Members that have access to a Master or Backup (or both), despite other link or Member failures, continue to operate as part of their active stack.

If multiple stack links or stack Member switches fail, and separate the Master and Backup into separate substacks, the Backup automatically becomes an active Master for the partial stack in which it is. Later, if the topology failures are corrected, the partial stacks merge, and the two active Masters come into contact.

In this scenario, if both the (original) Master and the Backup (acting as Master) are in operation when the merger occurs, the original Master reasserts its role as active Master for the entire stack. If any configuration elements are changed and applied on the Backup during the time it acted as the Master (and forwarded to its connected Members), the Backup and its affected Members reboot and are reconfigured by the returning Master before resuming their regular roles.

However, if the original Master switch is disrupted (powered down or in the process of rebooting) when it is reconnected with the active stack, the Backup (acting as Master) retains its acting Master status to avoid disruption to the functioning stack. The original Master temporarily assumes a role as a Backup.

If both the Master and Backup are rebooted, the switches assume their originally configured roles.

If, while the stack is still split, the Backup (acting as Master) is reconfigured to become a regular Master, then when the split stacks are finally merged, the Master with the lowest MAC address becomes the new active Master for the entire stack.

Merging independent stacks

If switches from different stacks are linked together in a stack topology without first reconfiguring their roles, it is possible that more than one switch in the stack might be configured as a Master.

Important: Although all switches that are configured for stacking and joined by stacking links are recognized as potential stack participants by any operational Master switches, they are not brought into operation within the stack until assigned (or “bound”) to a specific Master switch.

Consider two independent stacks, Stack A and Stack B, which are merged into one stacking topology. The stacks behave independently until the switches in Stack B are bound to Master A (or vice versa). In this example, after the Stack B switches are bound to Master A, Master A automatically reconfigures them to operate as Stack A Members, regardless of their original status within Stack B.

However, for future Backup selection purposes, reconfigured Masters retain their identity as configured Masters, even though they otherwise act as Members and lose all settings that pertain to their original stacks.

Backup switch selection

An operational stack can have one optional Backup at any time. Only the Backup specified in the active Master’s configuration is eligible to take over current stack control when the Master is rebooted or fails. The Master automatically synchronizes configuration settings with the specified Backup to facilitate the transfer of control functions.

The Backup retains its status until one of the following occurs:

- The Backup setting is deleted or changed from the active Master.
- A new Master assumes operation as the active Master in the stack, and uses its own configured Backup settings.

- The active Master is rebooted with the boot configuration set to factory defaults (clearing the Backup setting).

Master failover

When the Master switch is present, it controls the operation of the stack and pushes configuration information to the other switches in the stack. If the active Master fails, then the designated Backup (if one is defined in the Master's configuration) becomes the new acting Master and the stack continues to operate normally.

Secondary backup

When a Backup takes over stack control operations, if any other configured Masters (acting as Member switches) are available within the stack, the Backup selects one as a secondary Backup. The primary Backup automatically reconfigures the secondary Backup and specifies itself (the primary Backup) as the new Backup in case the secondary fails. This action prevents the chain of stack control from moving too far from the original Master and Backup configuration intended by the administrator.

Master recovery

If the prior Master recovers in a functioning stack where the Backup assumed stack control, the prior Master does not reassert itself as the stack Master. Instead, the prior Master assumes a role as a secondary Backup to avoid further stack disruption.

Upon stack reboot, the Master and Backup resume their regular roles.

No backup

If a Backup is not configured on the active Master, or the specified Backup is not operating, then if the active Master fails, the stack reboots without an active Master.

When a group of stacked switches are rebooted without an active Master present, the switches are considered to be isolated. All isolated switches in the stack are placed in a WAITING state until a Master appears. During this WAITING period, all the external ports and internal server ports of these Member switches are placed into operator-disabled state. Without the Master, a stack cannot respond correctly to networking events.

Stack Member identification

Each switch in the stack has two numeric identifiers, as follows:

- Attached Switch Number (asnum): An asnum is automatically assigned by the Master switch, based on each Member switch's physical connection in relation to the Master. The asnum is used as an internal ID by the Master switch and is not user-configurable.
- Configured Switch Number (csnum): The csnum is the logical switch ID assigned by the stack administrator. The csnum is used in most stacking-related configuration commands and switch information output. It is also used as a port prefix to distinguish the relationship between the ports on different switches in the stack.

You should use asnum 1 and csnum 1 to identify the Master switch. By default, csnum 1 is assigned to the Master. If csnum 1 is not available, the lowest available csnum is assigned to the Master.

Configuring a stack

This section provides procedures for creating a stack of switches. The high-level procedure is as follows:

1. Choose one Master switch for the entire stack.
2. Set all stack switches to stacking mode.

3. Configure the same stacking VLAN for all switches in the stack.
4. Configure the stacking interlinks.
5. Configure an external IP interface on the Master (if external management is wanted).
6. Bind Member switches to the Master.
7. Assign a Backup switch.

These tasks are covered in detail in the following sections.

Preferred configuration practices

Here are guidelines for building an effective switch stack:

- Always connect the stack switches in a complete ring topology.
- Avoid disrupting the stack connections unnecessarily while the stack is in operation.
- For enhanced redundancy when creating port trunks, include ports from different stack members in the trunks.
- Avoid altering the stack asnum and csnum definitions unnecessarily while the stack is in operation.
- When in stacking mode, the highest QoS priority queue is reserved for internal stacking requirements. Therefore, only seven priority queues are available for regular QoS use.
- Configure only as many QoS levels as necessary. This action allows for the best usage of packet buffers.

Configuring each switch in a stack

To configure each switch for stacking, connect to the internal management IP interface for each switch (assigned by the management system) and use the CLI to perform the following steps.

Important: Stacking configuration is stored as boot parameters. It is not part of the running configuration. Restoring the switch to its default configuration file does not delete the stacking parameters.

1. Enable stacking.

   On each switch, enable stacking by running **`boot stack enable`** (Example 6-1).

   Example 6-1 Enable stacking

```
ACC-3#boot stack enable
Current status: disabled
New status:     enabled
Next boot will have stacking enabled instead of disabled.
ACC-3#

ACC-4#boot stack enable
Current status: disabled
New status:     enabled
Next boot will have stacking enabled instead of disabled.
ACC-4#
```

2. Set the stacking membership mode.

   On each switch, set the stacking membership mode by running **boot stack mode {master|member}** (Example 6-2).

   By default, each switch is set to Member mode. However, one switch must be set to Master mode.

Example 6-2 Set stacking membership mode

```
ACC-3#boot stack mode master
Current switch stacking mode: Member
New switch stacking mode    : Master
A Reboot is required for the new settings to take effect
ACC-3#

ACC-4#boot stack mode member
Current switch stacking mode: Member
New switch stacking mode    : Member
ACC-4#
```

3. Configure the stacking VLAN.

   On each switch, configure the stacking VLAN (or use the default setting) by running **boot stack vlan** (Example 6-3).

   Important: Although any VLAN (except VLAN 1) may be defined for stack traffic, it is preferable that the default, VLAN 4090, be reserved for stacking, as shown in Example 6-3.

Example 6-3 Configure stacking VLAN

```
ACC-3#boot stack vlan 4090

ACC-4#boot stack vlan 4090
```

4. Designate the stacking links

   To create the preferred topology, at least two 10 Gb external ports on each switch should be dedicated to stacking.

   Important: By default, the 10 Gb Ethernet ports EXT1 and EXT2 are used.

   Run **boot stack higig-trunk <list of port names or aliases>** to specify the links to be used in the stacking trunk (Example 6-4). In the reference architecture, ports EXT9 and EXT10 are used for stacking.

Example 6-4 Designate stacking links

```
ACC-3#boot stack higig-trunk EXT9,EXT10

ACC-4#boot stack higig-trunk EXT9,EXT10
```

5. Physically connect the stack trunks.

   Connect the stacking links (Figure 6-1) and verify that the link is up.

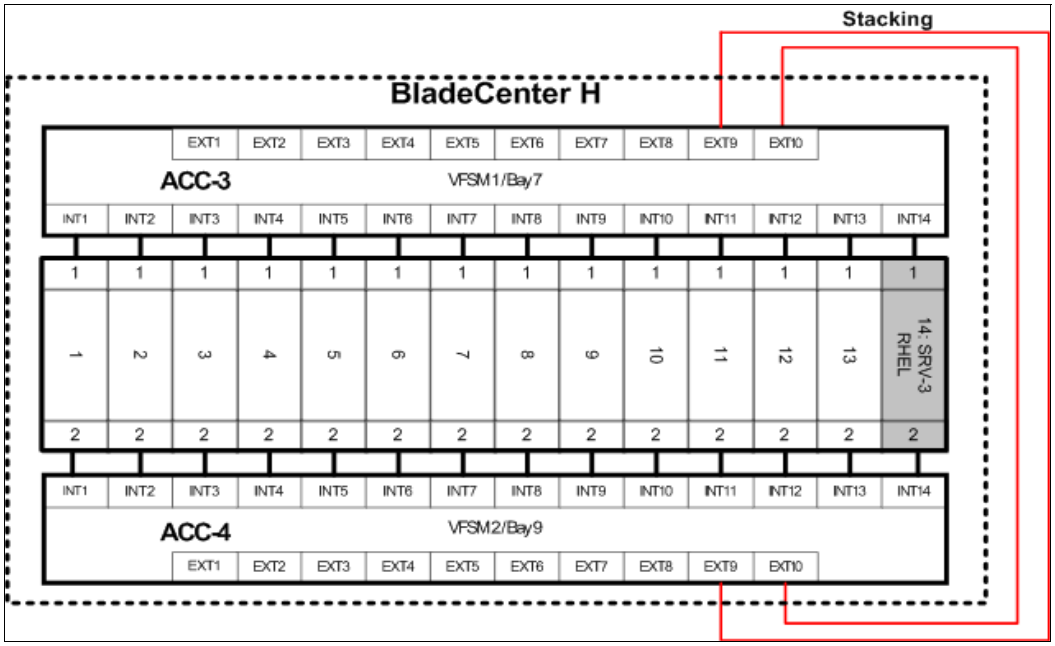
Stacking
BladeCenter H
EXT1 EXT2 EXT3 EXT4 EXT5 EXT6 EXT7 EXT8 EXT9 EXT10
ACC-3
VFSM1/Bay7
INT1 INT2 INT3 INT4 INT5 INT6 INT7 INT8 INT9 INT10 INT11 INT12 INT13 INT14
1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 2 3 4 5 6 7 8 9 10 11 12 13 14:SRV-3 RHEL
2 2 2 2 2 2 2 2 2 2 2 2 2 2
INT1 INT2 INT3 INT4 INT5 INT6 INT7 INT8 INT9 INT10 INT11 INT12 INT13 INT14
ACC-4
VFSM2/Bay9
EXT1 EXT2 EXT3 EXT4 EXT5 EXT6 EXT7 EXT8 EXT9 EXT10

Figure 6-1 Stack links physical connections

6. Verify the stacking configuration.

   Verify the saved stacking configuration by running **`show boot stack`** (Example 6-5).

Example 6-5 Verify the stacking configuration

```
ACC-3#show boot stack
Current stacking settings:
Stacking         : OFF
Switch Mode      : n/a
Stack Trunk Ports: empty
Stack VLAN       : 4090

Stacking settings saved:
Stacking         : ON
Switch Mode      : Master
Stack Trunk Ports: EXT9 EXT10
Stack VLAN       : 4090
ACC-3#


ACC-4#show boot stack
Current stacking settings:
Stacking         : OFF
Switch Mode      : n/a
Stack Trunk Ports: empty
Stack VLAN       : 4090

Stacking settings saved:
Stacking         : ON
```

```
Switch Mode      : Member
Stack Trunk Ports: EXT9 EXT10
Stack VLAN       : 4090
ACC-4#
```

7. Reboot the switches and verify the stack operation.

Important: After the switches restart, the stack is formed but it is not yet operational. Some observations can be made:

- The switches loaded the default configuration (see Example 6-6).
- The member switch is attached to the stack but is not bound to it (see Example 6-6).
- The member switches interfaces are not operational. Their link status is `dettached` (see Example 6-7).
- The interface status-related commands show all 216 ports (eight switches x 27 ports) for all eight maximum stack members, regardless if they exist or not (see Example 6-7).

Reboot the switches and run **`show stack switch`** to verify the current stack status (Example 6-6).

Example 6-6 Verify the stack status

```
Router#show stack switch
Stack name:
Local switch is the master.

Local switch:
   csnum             - 1
   MAC               - 00:25:03:6e:77:00
   UUID              - 05e9050bcd92450f903d7e60c581e4a4
   Bay Number        - 7
   Switch Type       - 7 (1-10GbESM)
   Chassis Type      - 2 (BladeCenter H)
   Switch Mode (cfg) - Master
   Priority          - 225
   Stack MAC         - 00:25:03:6e:77:1f

Master switch:
   csnum             - 1
   MAC               - 00:25:03:6e:77:00
   UUID              - 05e9050bcd92450f903d7e60c581e4a4
Press q to quit, any other key to continue
Bay Number        - 7

Backup switch:      not learnt yet.

Configured Switches:
------------------------------------------------------------------------
csnum             UUID                 Bay       MAC         asnum
------------------------------------------------------------------------
 C1   05e9050bcd92450f903d7e60c581e4a4  7  00:25:03:6e:77:00  A1

Attached Switches in Stack:
------------------------------------------------------------------------
```

```
asnum             UUID                  Bay       MAC          csnum  State
-------------------------------------------------------------------------------
 A1   05e9050bcd92450f903d7e60c581e4a4  7  00:25:03:6e:77:00  C1   IN_STACK
 A2   05e9050bcd92450f903d7e60c581e4a4  9  fc:cf:62:0a:49:00       ATTACH

Router#
```

Run **show interface status** to verify the stack members port status (Example 6-7).

Example 6-7 Stack members port status

```
Router#show interface status
------------------------------------------------------------------
Alias   Port   Speed    Duplex     Flow Ctrl      Link
------- ----   -----   --------  --TX-----RX--   ------
1:1     1      1G/10G    full     yes    yes      down
1:2     2      10000     full      no     no       up
1:3     3      10000     full      no     no       up
1:4     4      10000     full      no     no       up
1:5     5      10000     full      no     no       up
1:6     6      1G/10G    full     yes    yes      down
1:7     7      1G/10G    full     yes    yes      down
1:8     8      1G/10G    full     yes    yes      down
1:9     9      1G/10G    full     yes    yes      down
1:10    10     1G/10G    full     yes    yes      down
1:11    11     1G/10G    full     yes    yes      down
1:12    12     1G/10G    full     yes    yes      down
1:13    13     1G/10G    full     yes    yes      down
1:14    14     10000     full      no     no       up
1:15    15       100     full     yes    yes      up
1:16    16       100     full     yes    yes      down
1:17    17     10000     full      no     no      down
1:18    18     10000     full      no     no      down
1:19    19     1G/10G    full      no     no    disabled
1:20    20     10000     full      no     no       up
1:21    21      1000     full      no     no      down
1:22    22     1G/10G    full      no     no    disabled
1:23    23     1G/10G    full      no     no    disabled
1:24    24     1G/10G    full      no     no    disabled
1:25    25     10000     full      no     no       up
1:26    26     10000     full      no     no       up
1:27    27       any      any      no     no      down
2:1     65     1G/10G    full     yes    yes    detached
2:2     66     1G/10G    full     yes    yes    detached
2:3     67     1G/10G    full     yes    yes    detached
2:4     68     1G/10G    full     yes    yes    detached
2:5     69     1G/10G    full     yes    yes    detached
2:6     70     1G/10G    full     yes    yes    detached
2:7     71     1G/10G    full     yes    yes    detached
2:8     72     1G/10G    full     yes    yes    detached
2:9     73     1G/10G    full     yes    yes    detached
2:10    74     1G/10G    full     yes    yes    detached
2:11    75     1G/10G    full     yes    yes    detached
2:12    76     1G/10G    full     yes    yes    detached
2:13    77     1G/10G    full     yes    yes    detached
2:14    78     1G/10G    full     yes    yes    detached
```

```
2:15    79       100     full     yes    yes    detached
2:16    80       100     full     yes    yes    detached
2:17    81     1G/10G    full      no     no    detached
2:18    82     1G/10G    full      no     no    detached
2:19    83     1G/10G    full      no     no    detached
2:20    84     1G/10G    full      no     no    detached
2:21    85     1G/10G    full      no     no    detached
2:22    86     1G/10G    full      no     no    detached
2:23    87     1G/10G    full      no     no    detached
2:24    88     1G/10G    full      no     no    detached
2:25    89     1G/10G    full      no     no    detached
2:26    90     1G/10G    full      no     no    detached
2:27    91       any      any      no     no    detached
3:1     129    1G/10G    full     yes    yes    detached
3:2     130    1G/10G    full     yes    yes    detached
3:3     131    1G/10G    full     yes    yes    detached
3:4     132    1G/10G    full     yes    yes    detached
3:5     133    1G/10G    full     yes    yes    detached
3:6     134    1G/10G    full     yes    yes    detached
3:7     135    1G/10G    full     yes    yes    detached
3:8     136    1G/10G    full     yes    yes    detached
3:9     137    1G/10G    full     yes    yes    detached
3:10    138    1G/10G    full     yes    yes    detached
3:11    139    1G/10G    full     yes    yes    detached
3:12    140    1G/10G    full     yes    yes    detached
3:13    141    1G/10G    full     yes    yes    detached
3:14    142    1G/10G    full     yes    yes    detached
3:15    143      100     full     yes    yes    detached
3:16    144      100     full     yes    yes    detached
3:17    145    1G/10G    full      no     no    detached
3:18    146    1G/10G    full      no     no    detached
3:19    147    1G/10G    full      no     no    detached
3:20    148    1G/10G    full      no     no    detached
3:21    149    1G/10G    full      no     no    detached
3:22    150    1G/10G    full      no     no    detached
3:23    151    1G/10G    full      no     no    detached
3:24    152    1G/10G    full      no     no    detached
3:25    153    1G/10G    full      no     no    detached
3:26    154    1G/10G    full      no     no    detached
3:27    155      any      any      no     no    detached
4:1     193    1G/10G    full     yes    yes    detached
4:2     194    1G/10G    full     yes    yes    detached
4:3     195    1G/10G    full     yes    yes    detached
4:4     196    1G/10G    full     yes    yes    detached
4:5     197    1G/10G    full     yes    yes    detached
4:6     198    1G/10G    full     yes    yes    detached
4:7     199    1G/10G    full     yes    yes    detached
4:8     200    1G/10G    full     yes    yes    detached
4:9     201    1G/10G    full     yes    yes    detached
4:10    202    1G/10G    full     yes    yes    detached
4:11    203    1G/10G    full     yes    yes    detached
4:12    204    1G/10G    full     yes    yes    detached
4:13    205    1G/10G    full     yes    yes    detached
4:14    206    1G/10G    full     yes    yes    detached
4:15    207      100     full     yes    yes    detached
```

```
4:16    208      100     full     yes     yes    detached
4:17    209    1G/10G    full      no      no    detached
4:18    210    1G/10G    full      no      no    detached
4:19    211    1G/10G    full      no      no    detached
4:20    212    1G/10G    full      no      no    detached
4:21    213    1G/10G    full      no      no    detached
4:22    214    1G/10G    full      no      no    detached
4:23    215    1G/10G    full      no      no    detached
4:24    216    1G/10G    full      no      no    detached
4:25    217    1G/10G    full      no      no    detached
4:26    218    1G/10G    full      no      no    detached
4:27    219      any      any      no      no    detached
5:1     257    1G/10G    full     yes     yes    detached
5:2     258    1G/10G    full     yes     yes    detached
5:3     259    1G/10G    full     yes     yes    detached
5:4     260    1G/10G    full     yes     yes    detached
5:5     261    1G/10G    full     yes     yes    detached
5:6     262    1G/10G    full     yes     yes    detached
5:7     263    1G/10G    full     yes     yes    detached
5:8     264    1G/10G    full     yes     yes    detached
5:9     265    1G/10G    full     yes     yes    detached
5:10    266    1G/10G    full     yes     yes    detached
5:11    267    1G/10G    full     yes     yes    detached
5:12    268    1G/10G    full     yes     yes    detached
5:13    269    1G/10G    full     yes     yes    detached
5:14    270    1G/10G    full     yes     yes    detached
5:15    271      100     full     yes     yes    detached
5:16    272      100     full     yes     yes    detached
5:17    273    1G/10G    full      no      no    detached
5:18    274    1G/10G    full      no      no    detached
5:19    275    1G/10G    full      no      no    detached
5:20    276    1G/10G    full      no      no    detached
5:21    277    1G/10G    full      no      no    detached
5:22    278    1G/10G    full      no      no    detached
5:23    279    1G/10G    full      no      no    detached
5:24    280    1G/10G    full      no      no    detached
5:25    281    1G/10G    full      no      no    detached
5:26    282    1G/10G    full      no      no    detached
5:27    283      any      any      no      no    detached
6:1     321    1G/10G    full     yes     yes    detached
6:2     322    1G/10G    full     yes     yes    detached
6:3     323    1G/10G    full     yes     yes    detached
6:4     324    1G/10G    full     yes     yes    detached
6:5     325    1G/10G    full     yes     yes    detached
6:6     326    1G/10G    full     yes     yes    detached
6:7     327    1G/10G    full     yes     yes    detached
6:8     328    1G/10G    full     yes     yes    detached
6:9     329    1G/10G    full     yes     yes    detached
6:10    330    1G/10G    full     yes     yes    detached
6:11    331    1G/10G    full     yes     yes    detached
6:12    332    1G/10G    full     yes     yes    detached
6:13    333    1G/10G    full     yes     yes    detached
6:14    334    1G/10G    full     yes     yes    detached
6:15    335      100     full     yes     yes    detached
6:16    336      100     full     yes     yes    detached
```

```
6:17    337    1G/10G    full      no     no    detached
6:18    338    1G/10G    full      no     no    detached
6:19    339    1G/10G    full      no     no    detached
6:20    340    1G/10G    full      no     no    detached
6:21    341    1G/10G    full      no     no    detached
6:22    342    1G/10G    full      no     no    detached
6:23    343    1G/10G    full      no     no    detached
6:24    344    1G/10G    full      no     no    detached
6:25    345    1G/10G    full      no     no    detached
6:26    346    1G/10G    full      no     no    detached
6:27    347      any      any      no     no    detached
7:1     385    1G/10G    full     yes    yes    detached
7:2     386    1G/10G    full     yes    yes    detached
7:3     387    1G/10G    full     yes    yes    detached
7:4     388    1G/10G    full     yes    yes    detached
7:5     389    1G/10G    full     yes    yes    detached
7:6     390    1G/10G    full     yes    yes    detached
7:7     391    1G/10G    full     yes    yes    detached
7:8     392    1G/10G    full     yes    yes   detached
7:9     393    1G/10G    full     yes    yes    detached
7:10    394    1G/10G    full     yes    yes    detached
7:11    395    1G/10G    full     yes    yes    detached
7:12    396    1G/10G    full     yes    yes    detached
7:13    397    1G/10G    full     yes    yes    detached
7:14    398    1G/10G    full     yes    yes    detached
7:15    399      100     full     yes    yes    detached
7:16    400      100     full     yes    yes    detached
7:17    401    1G/10G    full      no     no    detached
7:18    402    1G/10G    full      no     no    detached
7:19    403    1G/10G    full      no     no    detached
7:20    404    1G/10G    full      no     no    detached
7:21    405    1G/10G    full      no     no    detached
7:22    406    1G/10G    full      no     no    detached
7:23    407    1G/10G    full      no     no    detached
7:24    408    1G/10G    full      no     no    detached
7:25    409    1G/10G    full      no     no    detached
7:26    410    1G/10G    full      no     no    detached
7:27    411      any      any      no     no    detached
8:1     449    1G/10G    full     yes    yes    detached
8:2     450    1G/10G    full     yes    yes    detached
8:3     451    1G/10G    full     yes    yes    detached
8:4     452    1G/10G    full     yes    yes    detached
8:5     453    1G/10G    full     yes    yes    detached
8:6     454    1G/10G    full     yes    yes    detached
8:7     455    1G/10G    full     yes    yes    detached
8:8     456    1G/10G    full     yes    yes    detached
8:9     457    1G/10G    full     yes    yes    detached
8:10    458    1G/10G    full     yes    yes    detached
8:11    459    1G/10G    full     yes    yes    detached
8:12    460    1G/10G    full     yes    yes    detached
8:13    461    1G/10G    full     yes    yes    detached
8:14    462    1G/10G    full     yes    yes    detached
8:15    463      100     full     yes    yes    detached
8:16    464      100     full     yes    yes    detached
8:17    465    1G/10G    full      no     no    detached
```

```
8:18    466    1G/10G    full      no     no    detached
8:19    467    1G/10G    full      no     no    detached
8:20    468    1G/10G    full      no     no    detached
8:21    469    1G/10G    full      no     no    detached
8:22    470    1G/10G    full      no     no    detached
8:23    471    1G/10G    full      no     no    detached
8:24    472    1G/10G    full      no     no    detached
8:25    473    1G/10G    full      no     no    detached
8:26    474    1G/10G    full      no     no    detached
8:27    475      any      any      no     no    detached
Router#
```

8. Bind Members to the stack.

Important: To fully activate the Member switches, you must bind them to the stack.

You can bind Member switches to stack *csnum* by using either their *asnum* or their *chassis UUID* and *bay number*.

Run **`stack switch-number <switch number 1-8> bind <asnum 1-16>`** to bind a Member to the stack (Example 6-8).

Example 6-8 Bind a Member to the stack

```
Router#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
Router(config)#stack switch-number 2 bind 2
```

The Member switch joins the stack and the interfaces start. The Member switch status is now `IN_STACK` (Example 6-9).

Example 6-9 Bound stack Member

```
Router#show stack switch
Stack name:
Local switch is the master.

Local switch:
   csnum             - 1
   MAC               - 00:25:03:6e:77:00
   UUID              - 05e9050bcd92450f903d7e60c581e4a4
   Bay Number        - 7
   Switch Type       - 7 (1-10GbESM)
   Chassis Type      - 2 (BladeCenter H)
   Switch Mode (cfg) - Master
   Priority          - 225
   Stack MAC         - 00:25:03:6e:77:1f

Master switch:
   csnum             - 1
   MAC               - 00:25:03:6e:77:00
   UUID              - 05e9050bcd92450f903d7e60c581e4a4
   Bay Number        - 7

Backup switch:      not learnt yet.

Configured Switches:
```

```
--------------------------------------------------------------------
csnum             UUID                  Bay       MAC         asnum
--------------------------------------------------------------------
 C1   05e9050bcd92450f903d7e60c581e4a4  7  00:25:03:6e:77:00  A1
 C2   05e9050bcd92450f903d7e60c581e4a4  9  fc:cf:62:0a:49:00  A2

Attached Switches in Stack:
-------------------------------------------------------------------------------
asnum             UUID                  Bay       MAC         csnum  State
-------------------------------------------------------------------------------
 A1   05e9050bcd92450f903d7e60c581e4a4  7  00:25:03:6e:77:00  C1   IN_STACK
 A2   05e9050bcd92450f903d7e60c581e4a4  9  fc:cf:62:0a:49:00  C2   IN_STACK

Router#
```

9. Assign a Backup switch (optional).

   To define a Member switch as a Backup that assumes the Master role if the Master switch should fail, run **stack backup <1-8>** (Example 6-10).

Example 6-10 Assign a Backup switch

```
Router#
Router#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
Router(config)#stack backup 2

Aug 17  6:59:27 ACC-3 NOTICE  stacking: BE_BACKUP sent to fc:cf:62:0a:49:00

Aug 17  6:59:27 ACC-3 NOTICE  stacking: Unit cpu fc:cf:62:0a:49:00 : BE_BACKUP
received from the master 00:25:03:6e:77:00

Aug 17  6:59:27 ACC-3 NOTICE  stacking: I_AM_BACKUP received from
fc:cf:62:0a:49:00

Router(config)#^Z
Router#
```

The **show stack switch** command shows the Backup switch information (Example 6-11).

Example 6-11 Configured Backup switch

```
Router#show stack switch
Stack name:
Local switch is the master.

Local switch:
   csnum             - 1
   MAC               - 00:25:03:6e:77:00
   UUID              - 05e9050bcd92450f903d7e60c581e4a4
   Bay Number        - 7
   Switch Type       - 7 (1-10GbESM)
   Chassis Type      - 2 (BladeCenter H)
   Switch Mode (cfg) - Master
   Priority          - 225
   Stack MAC         - 00:25:03:6e:77:1f
```

```
Master switch:
   csnum            - 1
   MAC              - 00:25:03:6e:77:00
   UUID             - 05e9050bcd92450f903d7e60c581e4a4
   Bay Number       - 7

Backup switch:
   csnum            - 2
   MAC              - fc:cf:62:0a:49:00
   UUID             - 05e9050bcd92450f903d7e60c581e4a4
   Bay Number       - 9

Configured Switches:
------------------------------------------------------------------------
csnum            UUID                 Bay       MAC         asnum
------------------------------------------------------------------------
 C1   05e9050bcd92450f903d7e60c581e4a4  7  00:25:03:6e:77:00  A1
 C2   05e9050bcd92450f903d7e60c581e4a4  9  fc:cf:62:0a:49:00  A2

Attached Switches in Stack:
------------------------------------------------------------------------------
asnum            UUID                 Bay       MAC         csnum  State
------------------------------------------------------------------------------
 A1   05e9050bcd92450f903d7e60c581e4a4  7  00:25:03:6e:77:00  C1   IN_STACK
 A2   05e9050bcd92450f903d7e60c581e4a4  9  fc:cf:62:0a:49:00  C2   IN_STACK

Router#
```

Important: The stack is now fully operational. You can now configure other features.

For more details about managing and operating the stack, see 6.7, “More information” on page 284.

6.3 Layer 1 implementation

This chapter presents Layer 1 related configuration and verification information for the implemented reference architecture.

This chapter includes the following topics:

- The network topology for a Layer 1 configuration
- The configuration of the port settings

6.3.1 Network topology for Layer 1 configuration

This section presents the Layer 1 implementation of the reference architecture. Figure 6-2shows the physical connections of the lab equipment that is used to demonstrate the examples in this chapter.

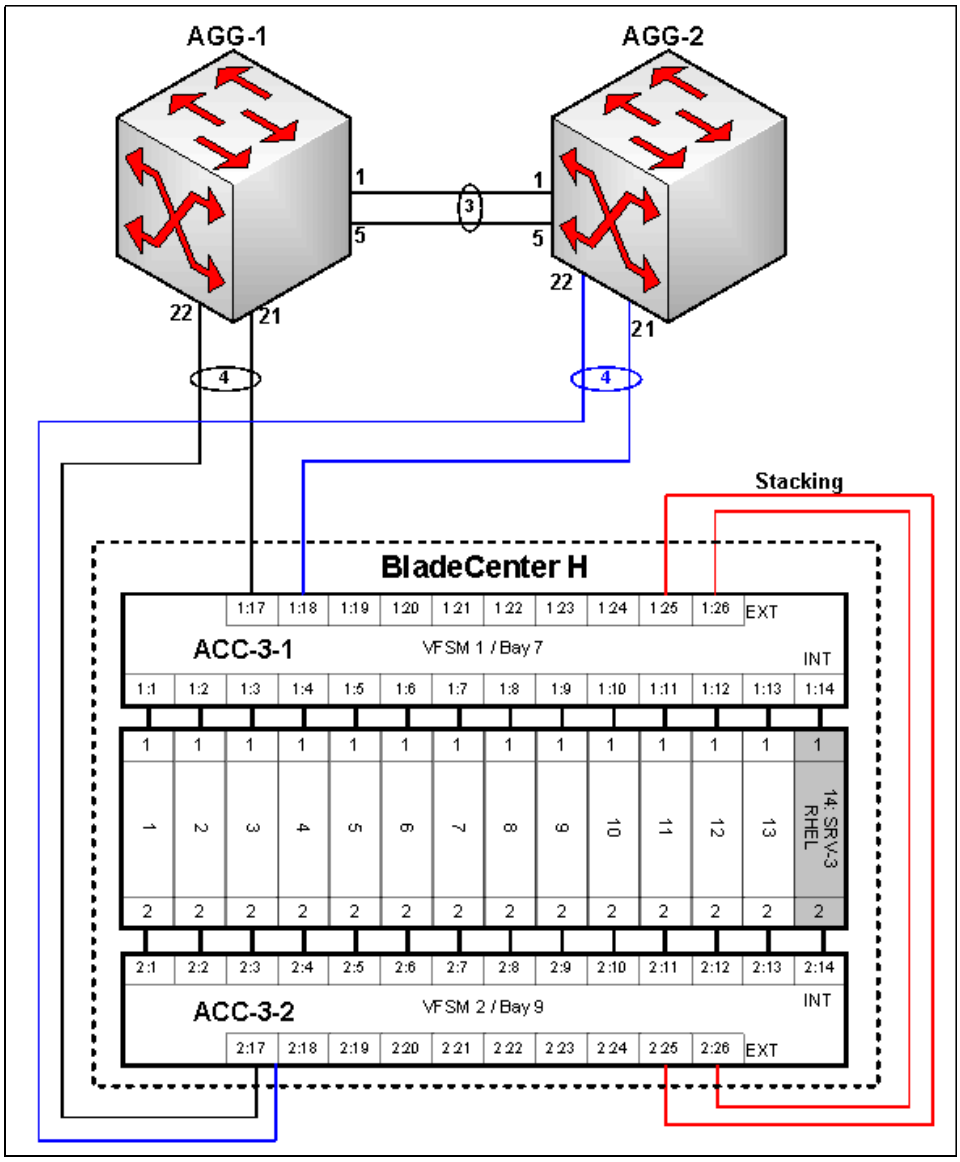

Figure 6-2 Physical topology

All the equipment used in the reference architecture uses IBM Networking OS V6.8, and the configuration, statistics, and information commands were tested on the switches.

6.3.2 Port settings configuration

The physical connection details are provided in Chapter 3, “Reference architectures” on page 107and most Layer 1 aspects do not involve any configuration. This section provides some useful commands for ports verification, to make sure all the links are up and running before proceeding to upper layer configuration.

Additional commands and details for Layer 1 configuration can be found in the technical documentation listed in 6.7, “More information” on page 284

Port link configuration

IBM switches include a factory default configuration that enables interfaces with the following link settings:

- In the copper Gigabit Ethernet interfaces:
  - Auto-negotiation is set.
  - The speed for 10/100/1000 RJ45 (copper) Gigabit Ethernet interfaces is set to auto, so that the interface can operate at 10 Mbps, 100 Mbps, or 1 Gbps. The link operates at the highest possible speed, depending on the capabilities of the remote end. If the speed is manually set to 1 Gbps, the duplex operation is automatically set to full.
  - The duplex mode is set to auto;
  - The flow control is set to none;
- In the fiber Gigabit Ethernet interfaces:
  - No auto-negotiation is set.
  - The speed is set to 1 Gbps.
  - The duplex mode is set to full.
  - The flow control is set to none.
- In the fiber 10 Gigabit Ethernet interfaces:
  - No auto-negotiation is set.
  - The speed is set to 10 Gbps.
  - The duplex mode is set to full.
  - The flow control is set to none.

All the ports used in this implementation are 10 Gbps (both external and internal) and thus default to no auto-negotiation / full-duplex. There is no need for a speed or duplex configuration.

The port configuration-related commands are listed here. To change the default link parameters on Gigabit Ethernet interfaces, enter one of the following interface/portchannel configuration level commands:

- **Switch(config-if)#speed {10|100|1000|auto}**
- **Switch(config-if)#duplex {full|half|any}**
- **Switch(config-if)#[no] flowcontrol {receive|send|both}**
- **Switch(config-if)#[no] auto**

Important: The default interface parameters are not listed in the configuration.

To temporarily disable a port, run the following command:

```
Switch#interface port <Port alias or number> shutdown
```

Layer 1 verification

Run the commands described in this section to verify the Layer 1 operation of the switch, including installed transceivers, link status, packet counters, and speed duplex.

Run **show interface transceiver** to show the installed transceivers type, part number, serial number, laser type, and status. Example 6-12 shows the command's output.

Example 6-12 Verification of the installed transceivers

```
ACC-3#show interface transceiver
------------------------------------------------------------------------------------------------
   Port     Device  TXEna RXSig TXuW  RXuW  TXFlt Vendor           Serial           Approval
---------- ------- ----- ----- ----- ----- ----- ---------------- ---------------- -----------
1:17 - EXT1 SR SFP+ Ena   Link  571.1 637.6 none  Blade Network    AA1018A3R7E      Approved
1:18 - EXT2 SR SFP+ Ena   Link  572.9 568.1 none  Blade Network    AA1022A4H55      Approved
1:19 - EXT3 NO device
1:20 - EXT4 SR SFP+ Ena   Link  572.9 470.4 none  Blade Network    AA1022A4GGS      Approved
1:21 - EXT5 SX SFP  Ena   Down  303.0 668.7 none  Blade Network    BNTM0941CD       Approved
1:22 - EXT6 NO device
1:23 - EXT7 NO device
1:24 - EXT8 NO device
1:25 - EXT9 1m DAC  Ena   Link   N/A   N/A  none  BLADE NETWORKS   APF09450020579   Approved
1:26 -EXT10 1m DAC  Ena   Link   N/A   N/A  none  BLADE NETWORKS   APF09450020668   Approved
2:17 - EXT1 SR SFP+ Ena   Link  573.5 637.6 none  Blade Network    AA1022A4F5M      Approved
2:18 - EXT2 SR SFP+ Ena   Link  570.6 566.8 none  Blade Network    AD0850ER0CR      Approved
2:19 - EXT3 NO device
2:20 - EXT4 NO device
2:21 - EXT5 NO device
2:22 - EXT6 NO device
2:23 - EXT7 NO device
2:24 - EXT8 NO device
2:25 - EXT9 1m DAC  Ena   Link   N/A   N/A  none  BLADE NETWORKS   APF09450020579   Approved
2:26 -EXT10 1m DAC  Ena   Link   N/A   N/A  none  BLADE NETWORKS   APF09450020668   Approved
ACC-3#
```

Run **show interface status** or **show interface link** to show information about link, duplex, speed, and flow control. Example 6-13 shows the command's output.

Example 6-13 Interface status verification

```
ACC-3#show interface status
------------------------------------------------------------------
Alias   Port   Speed    Duplex     Flow Ctrl      Link
------- ----   -----   --------  --TX-----RX--   ------
1:1     1      1G/10G    full     yes    yes      down
1:2     2      10000     full      no     no       up
1:3     3      10000     full      no     no       up
1:4     4      10000     full      no     no       up
1:5     5      10000     full      no     no       up
1:6     6      1G/10G    full     yes    yes      down
1:7     7      1G/10G    full     yes    yes      down
1:8     8      1G/10G    full     yes    yes      down
1:9     9      1G/10G    full     yes    yes      down
1:10    10     1G/10G    full     yes    yes      down
1:11    11     1G/10G    full     yes    yes      down
1:12    12     1G/10G    full     yes    yes      down
1:13    13     1G/10G    full     yes    yes      down
1:14    14     1G/10G    full     yes    yes      down
1:15    15       100     full     yes    yes       up
1:16    16       100     full     yes    yes      down
1:17    17     10000     full      no     no       up
1:18    18     10000     full      no     no       up
1:19    19     1G/10G    full      no     no    disabled
1:20    20     10000     full      no     no       up
```

```
1:21    21      1000     full      no     no      down
1:22    22     1G/10G    full      no     no    disabled
1:23    23     1G/10G    full      no     no    disabled
1:24    24     1G/10G    full      no     no    disabled
1:25    25     10000     full      no     no      up
1:26    26     10000     full      no     no      up
1:27    27       any      any      no     no     down
2:1     65     1G/10G    full     yes    yes     down
2:2     66     10000     full      no     no      up
2:3     67     1G/10G    full     yes    yes     down
2:4     68     10000     full      no     no      up
2:5     69     10000     full      no     no      up
2:6     70     1G/10G    full     yes    yes     down
2:7     71     1G/10G    full     yes    yes     down
2:8     72     10000     full      no     no      up
2:9     73     1G/10G    full     yes    yes     down
2:10    74     1G/10G    full     yes    yes     down
2:11    75     1G/10G    full     yes    yes     down
2:12    76     1G/10G    full     yes    yes     down
2:13    77     1G/10G    full     yes    yes     down
2:14    78     1G/10G    full     yes    yes     down
2:15    79       100     full     yes    yes      up
2:16    80       100     full     yes    yes     down
2:17    81     10000     full      no     no      up
2:18    82     10000     full      no     no     down
2:19    83     1G/10G    full      no     no    disabled
2:20    84     1G/10G    full      no     no    disabled
2:21    85     1G/10G    full      no     no    disabled
2:22    86     1G/10G    full      no     no    disabled
2:23    87     1G/10G    full      no     no    disabled
2:24    88     1G/10G    full      no     no    disabled
2:25    89     10000     full      no     no      up
2:26    90     10000     full      no     no      up
2:27    91       any      any      no     no     down
ACC-3#
```

Run **show interface counters** to show traffic statistics for the switch ports. Example 6-14 shows the command's output.

Example 6-14 Interface traffic statistics

```
AGG-1#show interface counters
------------------------------------------------------------------
output omitted...
------------------------------------------------------------------
Interface statistics for port 1:17:
                    ifHCIn Counters       ifHCOut Counters
Octets:                    19757662              11831311
UcastPkts:                      380                  1209
BroadcastPkts:                    1                   381
MulticastPkts:               251638                127094
FlowCtrlPkts:                     0                     0
PriFlowCtrlPkts:                  0                     0
Discards:                         0                     0
Errors:                           0                     0
```

```
Ingress Discard reasons for port 1:17:
VLAN Discards:                      0
Empty Egress Portmap:               0
Filter Discards:                    0
Policy Discards:                    0
Non-Forwarding State:               0
IBP/CBP Discards:                   0

----------------------------------------------------------------
Interface statistics for port 1:18:
                     ifHCIn Counters       ifHCOut Counters
Octets:                     49414808               4537173
UcastPkts:                         0                     0
BroadcastPkts:                  1845                     0
MulticastPkts:                712449                 15594
FlowCtrlPkts:                      0                     0
PriFlowCtrlPkts:                   0                     0
Discards:                     462838                462644
Errors:                            0                     0

Ingress Discard reasons for port 1:18:
VLAN Discards:                     0
Empty Egress Portmap:         462840
Filter Discards:                   0
Policy Discards:                   0
Non-Forwarding State:         462840
IBP/CBP Discards:                  0

----------------------------------------------------------------
Press q to quit, any other key to continue...
```

6.4 Layer 2 implementation

This section refers to basic Layer 2 configuration for the Virtual Fabric 10Gb Switch Module.

All the configurations presented in this section are implemented by using IBM Networking OS V6.8 installed in the reference architecture switches. Configuration steps and examples are presented for each command on selected equipment. It is assumed that the template can be easily replicated for the remaining ports and equipment according to the reference architecture. For the architecture planning details, such as VLAN numbers and assignment, and trunks, see Chapter 3, “Reference architectures” on page 107.

Not all the Layer 2 functions available in IBM Networking OS V6.8 are covered in this section. For an extensive list of features and configuration guidelines, see the documentation links listed in 6.7, “More information” on page 284.

The following topics are described in this section:

- VLANs
- Ports and trunking
- Spanning Tree Protocol (STP)
- Quality of Service (QoS)

6.4.1 VLANs

The VLAN-related configuration applied to the reference architecture switches is described in Chapter 3, “Reference architectures” on page 107. The configuration topics described in this section are:

- VLANs and port VLAN ID numbers
- VLAN tagging
- Private VLANs

VLANs and port VLAN ID numbers

This section shows some basic switching configuration, such as configuring a VLAN, assigning a port to a VLAN, and configuring private VLANs.

The Virtual Fabric 10Gb Switch Module supports up to 1024 VLANs per switch. Even though the maximum number of VLANs supported at any time is 1024, each can be identified with any number 1 - 4094.

VLAN 1 is the default VLAN for the external ports and the internal blade ports, so configure only those ports that belong to other VLANs.

VLAN 4095 is reserved for use by the management network, which includes internal management ports (MGT1 and MGT2) and (by default) internal ports.

VLANs definition and port assignment configuration steps

To define the VLANs and assign the ports, complete the following steps:

1. Define the VLANs by running the commands shown in Example 6-15.

   Example 6-15 VLAN definition

   ```
   ACC-3#configure terminal
   Enter configuration commands, one per line.  End with Ctrl/Z.
   ACC-3(config)#vlan 30

   VLAN number 30 with name "VLAN 30" created.

   VLAN 30 was assigned to STG 30.
   ACC-3(config-vlan)#name SRV-3
   ACC-3(config-vlan)#enable
   ACC-3(config-vlan)#^Z
   ACC-3#
   ```

2. Assign ports to a VLAN. If you want to assign a port to a VLAN, run **member** at the VLAN configuration level.

   Before you assign ports to VLANs, you should first define and name the VLANs. However, if you assign a port to a non-existent VLAN, the OS automatically creates one, according to the PVID defined for the port, and allocates a Spanning Tree Group for it. The VLAN port assignment configuration is shown in Example 6-16.

   Example 6-16 VLAN port assignment configuration

   ```
   ACC-3#configure terminal
   Enter configuration commands, one per line.  End with Ctrl/Z.
   ACC-3(config)#vlan 30
   ACC-3(config-vlan)#member 1:14,1:17,1:18,2:14,2:17,2:18
   Port 1:14 is an UNTAGGED port and its PVID is changed from 1 to 30
   ```

```
Port 1:17 is an UNTAGGED port and its PVID is changed from 1 to 30
Port 1:18 is an UNTAGGED port and its PVID is changed from 1 to 30
Port 2:14 is an UNTAGGED port and its PVID is changed from 1 to 30
Port 2:17 is an UNTAGGED port and its PVID is changed from 1 to 30
Port 2:18 is an UNTAGGED port and its PVID is changed from 1 to 30
ACC-3(config-vlan)#^Z
ACC-3#
```

3. To verify the VLANs, show the VLAN configuration on the switch by running **`show vlan [information]`**, as shown in Example 6-17.

Example 6-17 show vlan command output

```
ACC-3#show vlan
VLAN                Name                Status MGT           Ports
----  --------------------------------  ------ --- ------------------------
1     Default VLAN                      ena    dis  1:1-1:14 1:19-1:27 2:1-2:13
                                                     2:19-2:27 3:1-3:14
                                                     3:17-3:27 4:1-4:14
                                                     4:17-4:27 5:1-5:14
                                                     5:17-5:27 6:1-6:14
                                                     6:17-6:27 7:1-7:14
                                                     7:17-7:27 8:1-8:14
                                                     8:17-8:27
30    SRV-3                             ena    dis  1:14 1:17 1:18 2:14 2:17
                                                     2:18
4090  STK VLAN                          ena    dis  1:25 1:26 2:25 2:26
4095  Mgmt VLAN                         ena    ena  1:1-1:3 1:5-1:13 1:15 1:16
                                                     2:1-2:3 2:5-2:13 2:15 2:16
                                                     3:1-3:16 4:1-4:16 5:1-5:16
                                                     6:1-6:16 7:1-7:16 8:1-8:16
ACC-3#
```

Use the **`information`** option for a more detailed output.

VLAN tagging

IBM Networking OS supports 802.1Q VLAN tagging, providing standards-based VLAN support for Ethernet systems.

Important: The default configuration settings for VFSM switches are as follows:

- All external ports are untagged members of VLAN1 with PVID=1.
- All internal server ports are tagged members of VLAN1 and VLAN4095 with PVID=1.

For a detailed description of tagging and terminology, see 2.1, “Virtual Local Area Networks” on page 52.

Tagging is not used in our reference architecture for the links between ACC-3 and the aggregation switches, as only VLAN30 for SRV-3 needs to be carried to the aggregation layer. Trunks from ACC-3 to AGG-1, ACC-3 to AGG-2, and the host ports connected to SRV-3 are untagged members of VLAN30 with the PVID=30.

To enable tagging for an untagged port, run **tagging** in interface configuration mode (Example 6-18).

Example 6-18 Tagging configuration

```
ACC-3#configure terminal
ACC-3(config-if)#interface port 1:1
ACC-3(config-if)#tagging
ACC-3(config-if)#^Z
ACC-3#
```

To allow communication over a tagging enabled connection, the end ports of the switch must be declared members of the required VLANs to be transported over the link.

Information about tagging, port names, allocated VLANs, PVIDs, and other flags are summarized in the **show interface information** command output, as shown in Example 6-19.

Example 6-19 Display interface

```
ACC-3#show interface information
Alias   Port Tag    Type     RMON Lrn Fld PVID       NAME                 VLAN(s)
------- ---- --- ---------- ---- --- --- ----- -------------- --------------------
1:1     1     y  Internal    d    e   e     1  1:1            1 4095
1:2     2     y  Internal    d    e   e     1  1:2            1 4095
1:3     3     y  Internal    d    e   e     1  1:3            1 4095
1:4     4     y  Internal    d    e   e     1  1:4            1 4095
1:5     5     y  Internal    d    e   e     1  1:5            1 4095
1:6     6     y  Internal    d    e   e     1  1:6            1 4095
1:7     7     y  Internal    d    e   e     1  1:7            1 4095
1:8     8     y  Internal    d    e   e     1  1:8            1 4095
1:9     9     y  Internal    d    e   e     1  1:9            1 4095
1:10    10    y  Internal    d    e   e     1  1:10           1 4095
1:11    11    y  Internal    d    e   e     1  1:11           1 4095
1:12    12    y  Internal    d    e   e     1  1:12           1 4095
1:13    13    y  Internal    d    e   e     1  1:13           1 4095
1:14    14    y  Internal    d    e   e    30  SRV-3          30
1:15    15    y  LocalMgmt   d    e   e  4095* 1:15           4095
1:16    16    y  LocalMgmt   d    e   e  4095* 1:16           4095
1:17    17    n  External    d    e   e    30  AGG1-ACC3      30
1:18    18    n  External    d    e   e    30  AGG2-ACC3      30
1:19    19    n  External    d    e   e     1  1:19           1
1:20    20    n  External    d    e   e     1  1:20           1
1:21    21    n  External    d    e   e     1  1:21           1
1:22    22    n  External    d    e   e     1  1:22           1
1:23    23    n  External    d    e   e     1  1:23           1
1:24    24    n  External    d    e   e     1  1:24           1
1:25    25    n  Stacking    d    e   e     1  1:25           Stacking
1:26    26    n  Stacking    d    e   e     1  1:26           Stacking
1:27    27    n  External    d    e   e     1  1:27           1
2:1     65    y  Internal    d    e   e     1  2:1            1 4095
2:2     66    y  Internal    d    e   e     1  2:2            1 4095
2:3     67    y  Internal    d    e   e     1  2:3            1 4095
2:4     68    y  Internal    d    e   e     1  2:4            1 4095
2:5     69    y  Internal    d    e   e     1  2:5            1 4095
2:6     70    y  Internal    d    e   e     1  2:6            1 4095
```

```
2:7     71    y  Internal    d    e   e     1  2:7            1 4095
2:8     72    y  Internal    d    e   e     1  2:8            1 4095
2:9     73    y  Internal    d    e   e     1  2:9            1 4095
2:10    74    y  Internal    d    e   e     1  2:10           1 4095
2:11    75    y  Internal    d    e   e     1  2:11           1 4095
2:12    76    y  Internal    d    e   e     1  2:12           1 4095
2:13    77    y  Internal    d    e   e     1  2:13           1 4095
2:14    78    y  Internal    d    e   e    30  SRV-3          30
2:15    79    y  RemoteMgmt  d    e   e  4095* 2:15          4095
2:16    80    y  RemoteMgmt  d    e   e  4095* 2:16          4095
2:17    81    n  External    d    e   e    30  AGG1-ACC3      30
2:18    82    n  External    d    e   e    30  AGG2-ACC3      30
2:19    83    n  External    d    e   e     1  2:19           1
2:20    84    n  External    d    e   e     1  2:20           1
2:21    85    n  External    d    e   e     1  2:21           1
2:22    86    n  External    d    e   e     1  2:22           1
2:23    87    n  External    d    e   e     1  2:23           1
2:24    88    n  External    d    e   e     1  2:24           1
2:25    89    n  Stacking    d    e   e     1  2:25           Stacking
2:26    90    n  Stacking    d    e   e     1  2:26           Stacking
2:27    91    n  External    d    e   e     1  2:27           1
* = PVID is tagged.
ACC-3#
```

Private VLANs

This feature is not in the scope of the reference architecture implementation. However, a summary of commands used for configuration and verification is presented in this section.

For more information about private VLANs concept, see 2.1, “Virtual Local Area Networks” on page 52.

Use the following commands to configure Private VLANs:

- Run **`private-vlan type primary`** at the VLAN level to configure the VLAN type as a Primary VLAN. A Private VLAN must have only one primary VLAN. The primary VLAN carries unidirectional traffic to ports on the isolated VLAN or to community VLAN.
- Run **`private-vlan type community`** at the VLAN level to configure the VLAN type as a community VLAN. Community VLANs carry upstream traffic from host ports. A Private VLAN may have multiple community VLANs.
- Run **`private-vlan type isolated`** at the VLAN level to configure the VLAN type as an isolated VLAN. The isolated VLAN carries unidirectional traffic from host ports. A Private VLAN may have only one isolated VLAN.
- Run **`no private-vlan type`** to clear the private-VLAN type.
- Run **`[no] private-vlan map [<2-4094>]`** to configure Private VLAN mapping between a secondary VLAN and a primary VLAN. Enter the primary VLAN ID. Secondary VLANs have the type defined as isolated or community. Use the **`no`** option to remove the mapping between the secondary VLAN and the primary VLAN.
- Run **`[no] private-vlan enable`** at the VLAN level to enable or disable the Private VLAN.
- Run **`show private-vlan [<2-4094>]`** to show the current parameters for the selected Private VLANs.

6.4.2 Ports and trunking

When using port trunk groups between two switches, you can create a virtual link between the switches, operating with combined throughput levels that depend on how many physical ports are included.

Two trunk types are available: *Static trunk groups* (portchannel), and *dynamic LACP trunk groups*.

Up to 18 trunks of each type are supported in stand-alone (non-stacking) mode, and 64 trunks of each type are supported in stacking mode, depending of the number and type of available ports. Each trunk can include up to eight member ports.

Trunk groups are also useful for connecting a VFSM to third-party devices that support link aggregation, such as Cisco routers and switches with EtherChannel technology (not ISL trunking technology) and the Sun Quad Fast Ethernet Adapter. Trunk group technology is compatible with these devices when they are configured manually.

Trunk traffic is statistically distributed among the ports in a trunk group, based on various configurable options.

Also, because each trunk group is composed of multiple physical links, the trunk group is inherently fault tolerant. If one connection between the switches is available, the trunk remains active, and statistical load balancing is maintained whenever a port in a trunk group is lost or returned to service.

Static trunk group configuration rules

The trunking feature operates according to specific configuration rules. When creating trunks, consider the following rules that determine how a trunk group reacts in any network topology:

- All trunks must originate from one network entity (a single device, or multiple devices that act in a stack) and lead to one destination entity. For example, you cannot combine links from two different servers into one trunk group.
- Ports from different member switches in the same stack (see 6.2, “Stacking” on page 240) may be aggregated together in one trunk.
- Any physical switch port can belong to only one trunk group.
- Depending on port availability, the switch supports up to eight ports in each trunk group.
- Internal (INTx) and external ports (EXTx) cannot become members of the same trunk group.
- Trunking from third-party devices must comply with Cisco EtherChannel technology.
- All trunk member ports must be assigned to the same VLAN configuration before the trunk can be enabled.
- If you change the VLAN settings of any trunk member, you cannot apply the change until you change the VLAN settings of all trunk members.
- When an active port is configured in a trunk, the port becomes a trunk member when you enable the trunk by running **`/cfg/l2/trunk <x>/ena`**. The Spanning Tree parameters for the port then change to reflect the new trunk settings.
- All trunk members must be in the same Spanning Tree Group (STG) and can belong to only one STG. However if all ports are tagged, then all trunk ports can belong to multiple STGs.
- If you change the Spanning Tree participation of any trunk member to enabled or disabled, the Spanning Tree participation of all members of that trunk changes similarly.

- When a trunk is enabled, the trunk Spanning Tree participation setting takes precedence over any trunk member.
- 802.1X authentication is not supported on ISL ports or on any port that is part of a trunk.
- You cannot configure a trunk member as a monitor port in a port-mirroring configuration.
- Trunks cannot be monitored by a monitor port; however, trunk members can be monitored.
- All ports in static trunks must have the same link configuration (speed, duplex, and flow control).

Static trunks configuration

Note: Static trunks are configured in the reference architecture on the links between ACC-3 and AGG-1, and ACC-3 and AGG-2

By default, each trunk group is empty and disabled. To define a static trunk group, complete the following steps:

1. Add physical ports to a trunk group.

   Run **[no] portchannel <number> port <port alias or number>** from the global configuration mode to add or remove ports to or from a trunk group (Example 6-20).

   Example 6-20 Static trunk configuration

```
ACC-3#conf t
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-3(config)#portchannel 1 port 1:17,2:17
ACC-3(config)#portchannel 2 port 1:18,2:18
ACC-3(config)#^Z
ACC-3#
```

2. Enable the trunk group.

   Run **[no] portchannel <number> enable** from the global configuration mode to enable or disable trunk groups (Example 6-21).

   Example 6-21 Enable static trunks

```
ACC-3#conf t
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-3(config)#portchannel 1 enable
ACC-3(config)#portchannel 2 enable
ACC-3(config)#^Z
ACC-3#
```

   You can remove the current trunk group configuration by running **no portchannel <number>**.

3. Configure hashing.

   Traffic in a trunk group is statistically distributed among member ports by using a hash process where various address and attribute bits from each transmitted frame are recombined to specify the particular trunk port the frame uses.

   The switch can be configured to use various hashing options. To achieve the most even traffic distribution, select options that exhibit a wide range of values for your particular network. Avoid hashing of information that is not present in the expected traffic, or which does not vary.

Trunk hash parameters are set globally. You can enable one or two parameters to configure any of the following valid combinations:

- SMAC (source MAC only)
- DMAC (destination MAC only)
- SIP (source IP only)
- DIP (destination IP only)
- SIP + DIP (source IP and destination IP)
- SMAC + DMAC (source MAC and destination MAC)

For trunk hashing configuration commands, see Table 6-1.

Table 6-1 Trunk hashing configuration commands

Hash options	VFSM command
Layer 2	
SMAC	**portchannel thash l2thash l2-source-mac-address**
DMAC	**portchannel thash l2thash l2-destination-mac-address**
SMAC+DMAC	**portchannel thash l2thash l2-source-destination-mac**
Layer 3	
SIP	**portchannel thash l3thash l3-source-ip-address**
DIP	**portchannel thash l3thash l3-destination-ip-address**
SIP+DIP	**portchannel thash l3thash l3-source-destination-ip**
L3 use L2	**portchannel thash l3thash l3-use-l2-hash**
Other options	
Ingress port	**portchannel thash ingress**
L4 port	**portchannel thash L4port**

In our reference architecture, source-destination MAC address, source-destination IP address, ingress, and L4 port hashing are enabled by default. To configure the hashing options, run **portchannel thash** with the appropriate options (Example 6-22).

Example 6-22 Configure trunk hashing

```
ACC-3#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
ACC-3(config)#portchannel thash l2thash l2-source-mac-address
ACC-3(config)#portchannel thash ingress
ACC-3(config)#^Z
ACC-3#
```

Run **show portchannel hash** to verify the global hash parameters.

Example 6-23 Trunk hashing parameters

```
ACC-3#show portchannel hash
Current L2 trunk hash settings:
    smac dmac
Current L3 trunk hash settings:
    sip dip
Current ingress port hash: enabled
```

```
Current L4 port hash: enabled
ACC-3#
```

4. Verify the trunk group configuration.

   Run the following commands to verify the trunk configuration and status.

   To verify the trunk group status, run **show portchannel [information]** or **show portchannel <number> [information]** (Example 6-24).

Example 6-24 Trunk group status information

```
ACC-3#show portchannel information
PortChannel 1: Enabled
Protocol - Static
Port State:
   1:17: STG 30 forwarding
   2:17: STG 30 forwarding

PortChannel 2: Enabled
Protocol - Static
Port State:
   1:18: STG 30 BLOCKING
   2:18: STG 30 BLOCKING

ACC-3#
```

Important: You can see that the trunk that connects ACC-3 to AGG-2 is blocked by the STP.

Verify the trunk group parameters by running **show portchannel <number>** (Example 6-25).

Example 6-25 Trunk group parameters

```
ACC-3#show portchannel 1
Protocol - Static
Current settings: enabled
    ports:  1:17, 2:17
Current L2 trunk hash settings:
    smac dmac
Current L3 trunk hash settings:
    sip dip
Current ingress port hash: enabled
Current L4 port hash: enabled
ACC-3##
```

Dynamic trunks

LACP trunks are not configured in VFSM for our reference architecture. For detailed information about the configuration steps and verification, see Chapter 5, “IBM System Networking RackSwitch implementation” on page 155.

6.4.3 Spanning Tree Protocol

The STP used for the reference architecture is Per-VLAN Rapid Spanning Tree (PVRST).

PVRST mode is based on RSTP, which provides rapid Spanning Tree convergence, but allows for multiple Spanning Tree Groups (STGs), with STGs on a per-VLAN basis. PVRST mode is compatible with Cisco R-PVST/R-PVST+ mode.

To simplify the switch configuration, VLAN Automatic STG Assignment (VASA) can be used in SPT/PVST+ or PVRST modes. When VASA is enabled, it is no longer necessary to manually assign an STG for each new VLAN. Instead, each newly configured VLAN is automatically assigned its own STG. If an empty STG is not available, the VLAN is automatically assigned to the default VLAN. When a VLAN is deleted, if there is no other VLAN associated with the assigned STG, the STG is returned to the available pool.

Up to 128 STGs can be configured on the switch (STG 128 is reserved for management).

VASA is disabled by default, but can be enabled or disabled by running **spanning-tree stg-auto** as follows:

```
ACC-3(config)#spanning-tree stg-auto
Warning: all VLANs will be assinged to a STG automatically.
ACC-3(config)#
```

VASA applies only to STP/PVST+ and PVRST modes and is ignored in RSTP and MSTP modes. When VASA is enabled, manual STG assignment is still available. The administrator may assign a specific STG to a VLAN by using regular commands:

```
# spanning-tree stp <STG> vlan <vlan ID>
```

When changing to STP/PVST+ or PVSRT mode (either from RSTP or MSTP modes, or when STP is disabled), all existing VLANs are assigned to a unique STG.

For simplicity and the consistency of the numbering conventions of the reference architecture, the VLANs IDs are assigned numbers lower than 127 so that the VASA can also assign matching STG numbers for each VLAN (see Example 6-15 on page 261).

The STG configuration is shown in Example 6-26.

Example 6-26 VLANs and STGs configuration

```
vlan 30
        enable
        name "SRV-3"
        member 1:14,1:17-1:18
        member 2:14,2:17-2:18

spanning-tree stp 30 vlan 30
```

PVRST is configured in our reference architecture for VLAN30. The Spanning Tree configuration is consistent with the RackSwitch implementation in Chapter 5, “IBM System Networking RackSwitch implementation” on page 155.

The per-VLAN roles of the switches are:

AGG-1 (G8264) VLAN 30 PVRST primary root (priority 0)
AGG-2 (G8264) VLAN 30 PVRST secondary root (priority 4096)
ACC-3 (VFSM stack) VLAN 30 default PVRST priority 61440

Run **spanning-tree stp <STG number> bridge priority <0-65535>** to configure the bridge priority for VLAN 30, as shown in Example 6-27.

Example 6-27 Bridge priority configuration example

```
AGG-2#conf t
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-2(config)#spanning-tree stp 30 bridge priority 4096

Sep 19  1:35:02 AGG-2 ALERT   stg: STG 30, new root bridge

Sep 19  1:35:02 AGG-2 ALERT   stg: STG 30, topology change detected

AGG-2(config)#^Z

AGG-2#


AGG-1#conf t
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-1(config)#spanning-tree stp 30 bridge priority 0

Aug 17 10:04:48 AGG-1 ALERT   stg: STG 30, new root bridge

Aug 17 10:04:48 AGG-1 ALERT   stg: STG 30, topology change detected
AGG-1(config)#^Z

AGG-1#
```

The following commands appear under the Spanning Tree section of the configuration:

```
spanning-tree stp 30 bridge priority 0
spanning-tree stp 30 vlan 30

spanning-tree stp 30 bridge priority 4096
spanning-tree stp 30 vlan 30
```

For more detailed configuration options for Spanning Tree (global commands, timers, and ports parameters), see the guides referenced in 6.7, “More information” on page 284.

To verify the Spanning Tree configuration and operation, run **shown spanning-tree stp <STG number> [bridge|information]** for detailed information about STP operation. If you do not provide an STG number, the command output shows the STP information for all groups, as shown in Example 6-28. If you want to narrow the output to a specific STG, add the optional parameters.

Example 6-28 show spanning-tree command output

```
ACC-3#show spanning-tree stp 30
----------------------------------------------------------------
Current Spanning Tree Group 30 settings: ON (PVRST)

Bridge params:  Priority  Hello  MaxAge  FwdDel  Aging
                 61440      2      20      15     300
VLANs:  30
STP Ports:
Port 1:14 :  Priority 128, Path Cost 0, auto, edge, Spanning Tree turned OFF
```

```
Port 1:17 :  Priority 128, Path Cost 0, auto
Port 1:18 :  Priority 128, Path Cost 0, auto
Port 2:14 :  Priority 128, Path Cost 0, auto, edge, Spanning Tree turned OFF
Port 2:17 :  Priority 128, Path Cost 0, auto
Port 2:18 :  Priority 128, Path Cost 0, auto
ACC-3#
ACC-3#
ACC-3#show spanning-tree stp 30 information
----------------------------------------------------------------

----------------------------------------------------------------
Spanning Tree Group 30: On (PVRST)
VLANs:  30

Current Root:            Path-Cost  Port Hello MaxAge FwdDel
 001e 08:17:f4:32:c4:00      990    1:17    2     20     15

Parameters:  Priority  Hello  MaxAge  FwdDel  Aging   Topology Change Counts
              61470      2      20      15     300                   9

   Port       Prio   Cost     State  Role Designated Bridge      Des Port Type
------------- ---- ---------- -----  ---- ---------------------- -------- ----
1:14            0        0    DSB *
1:17   (pc1)   128      990!+  FWD   ROOT 001e-08:17:f4:32:c4:00     8045  P2P
1:18   (pc2)   128      990!+  DISC  ALTN 101e-fc:cf:62:9d:9a:00     8045  P2P
2:14            0        0    DSB *
2:17   (pc1)   128      990!+  FWD   ROOT 001e-08:17:f4:32:c4:00     8045  P2P
2:18   (pc2)   128      990!+  DISC  ALTN 101e-fc:cf:62:9d:9a:00     8045  P2P
* = STP turned off for this port.
! = Automatic path cost.
+ = Portchannel cost, not the individual port cost.
ACC-3#
```

6.4.4 Quality of Service

The Quality of Service (QoS) implementation in IBM Networking OS is not in the scope of the reference architecture used in this book. However, a summary of configuration topics and commands is presented here for completeness.

QoS commands configure the 802.1p priority value and DiffServ Code Point value of incoming packets. You can then differentiate between various types of traffic, and provide different priority levels.

802.1p configuration

This feature provides the switch the capability to filter IP packets based on the 802.1p bits in the packet's VLAN header. The 802.1p bits specify the priority that you should give to the packets while forwarding them. The packets with a higher (non-zero) priority bits are given forwarding preference over packets with numerically lower priority bits value.

To configure the 802.1p parameters, run the following commands:

- Run **qos transmit-queue mapping <priority (0-7)> <COSq number>** to map the 802.1p priority to the Class of Service queue (COSq) priority. Enter the 802.1p priority value (0 - 7), followed by the Class of Service queue that handles the matching traffic.
- Run **qos transmit-queue weight-cos <COSq number> <weight (0-15)>** to configure the weight of the selected Class of Service queue (COSq). Enter the queue number (0 - 1), followed by the scheduling weight (0 - 15).
- Run **show qos transmit-queue** to shows the current 802.1p parameters.

DSCP configuration

The following commands map the DiffServ Code Point (DSCP) value of incoming packets to a new value or to an 802.1p priority value:

- Run **qos dscp dscp-mapping <DSCP (0-63)> <new DSCP (0-63)>** to map the initial DiffServ Code Point (DSCP) value to a new value. Enter the DSCP value (0 - 63) of incoming packets, followed by the new value.
- Run **qos dscp dot1p-mapping <DSCP (0-63)> <priority (0-7)>** to map the DiffServ Code point value to an 802.1p priority value. Enter the DSCP value, followed by the corresponding 802.1p value.
- Run **[no] qos dscp re-marking** to turn on or off DSCP re-marking globally.
- Run **show qos dscp** to show the current DSCP parameters.

6.5 High availability

This section presents high availability mechanisms in the Virtual Fabric 10Gb Switch Module.

The topics described in this section are:

- Stacking
- Layer 2 Failover
- Trunking
- Hot Links
- VRRP

6.5.1 Stacking

A stack is a group of up to eight Virtual Fabric 10Gb Switch Module devices that work together as a unified system. Because the multiple members of a stack act as a single switch entity with distributed resources, high-availability topologies can be more easily achieved.

A simple stack using two switches provides full redundancy in the event that either switch fails.

Stacking permits ports within different physical switches to be trunked together, further enhancing switch redundancy.

For a detailed presentation of failover stacking mechanisms, see 6.2, “Stacking” on page 240.

6.5.2 Layer 2 Failover

The primary application for Layer 2 Failover is to support Network Adapter Teaming. With Network Adapter Teaming, all the NICs on each server share an IP address, and are configured into a team. One NIC is the primary link, and the other is a standby link. For more details, see the documentation for your Ethernet adapter.

Important: Only two links per server can be used for Layer 2 Trunk Failover (one primary and one backup). Network Adapter Teaming allows only one backup NIC for each server blade.

Layer 2 Failover can be enabled on any trunk group in the switch, including LACP trunks. Trunks can be added to failover trigger groups. Then, if some specified number of monitor links fail, the switch disables all the control ports in the switch. When the control ports are disabled, it causes the NIC team on the affected servers to fail over from the primary to the backup NIC. This process is called a *failover event*.

When the appropriate number of links in a monitor group return to service, the switch enables the control ports. This action causes the NIC team on the affected servers to fail back to the primary switch (unless Auto-Fallback is disabled on the NIC team). The backup switch processes traffic until the primary switch's control links come up, which can take up to 5 seconds.

For more details about the Layer 2 Failover feature, see Chapter 2, "IBM System Networking Switch 10Gb Ethernet switch features" on page 51.

Because the Virtual Fabric 10Gb Switch Modules are configured in stacking mode, they work together as a unified system and both blade server ports are practically connected to a single switch. The stacking operation already provides uplink redundancy.

Layer 2 Failover is most relevant when the VFSMs operate in stand-alone mode, with the blade server connected to each switch. For NIC teaming and Layer 2 Failover with stand-alone switches, see Chapter 2, "IBM System Networking Switch 10Gb Ethernet switch features" on page 51.

6.5.3 Trunking

Multiple switch ports can be combined together to form robust, high-bandwidth trunks to other devices. Since trunks are composed of multiple physical links, the trunk group is inherently fault tolerant. If one connection between the switches is available, the trunk remains active.

For detailed information about trunking, see 6.4.2, "Ports and trunking" on page 265.

6.5.4 Hot Links

Note: The Hot Links mechanism is not used in th reference architecture implementation. The following section is just a short outline of the feature. For more information, see the documentation listed in 6.7, "More information" on page 284.

For network topologies that require Spanning Tree to be turned off, Hot Links provides basic link redundancy with fast recovery.

Hot Links has up to 25 triggers. A trigger is a pair of Layer 2 interfaces, each containing an individual port, trunk, or LACP adminkey. One interface is the Master, and the other is a Backup. While the Master interface is set to the active state and forwards traffic, the Backup interface is set to the standby state and blocks traffic until the Master interface fails. If the Master interface fails, the Backup interface is set to active and forwards traffic. After the Master interface is restored, it moves to the standby state and blocks traffic until the Backup interface fails. You may select a physical port, static trunk, or an LACP adminkey as a Hot Link interface.

Configuration guidelines

The following configuration guidelines apply to Hot Links:

- Ports that are configured as Hot Link interfaces must have STP disabled.
- When Hot Links is turned on, MSTP, RSTP, and PVRST must be turned off.
- When Hot Links is turned on, UplinkFast must be disabled.
- A port that is a member of the Master interface cannot be a member of the Backup interface.
- A port that is a member of one Hot Links trigger cannot be a member of another Hot Links trigger.
- An individual port that is configured as a Hot Link interface cannot be a member of a trunk.

Configuring Hot Links

Run the following commands to configure Hot Links:

- Enable Hot Links Trigger 1 by running the following command:

```
Switch(config)# hotlinks trigger 1 enable
```

- Add a port to the Master interface by running the following command:

```
Switch(config)# hotlinks trigger 1 master port 1
```

- Add a port to a Backup interface by running the following command:

```
Switch(config)# hotlinks trigger 1 backup port 2
```

- Turn on Hot Links by running the following command:

```
Switch(config)# hotlinks enable
```

6.5.5 VRRP

The VRRP feature is not supported in VFSM stacking mode and is not needed because the stacking members act as a single environment.

For high availability at the aggregation layer, VRRP is implemented in AGG-1 and AGG-2 switches to provide gateway redundancy for SRV-3,

VRRP enables redundant router configurations within a LAN, providing alternative router paths for a host to eliminate single points of failure within a network. Each participating VRRP-capable routing device is configured with the same virtual router IPv4 address and ID number. One of the virtual routers is elected as the Master, based on a number of priority criteria, and assumes control of the shared virtual router IPv4 address. If the Master fails, one of the backup virtual routers takes control of the virtual router IPv4 address and actively process traffic addressed to it.

For detailed information about VRRP concepts and components, see the 2.7.6, “Virtual Router Redundancy Protocol” on page 82.

To enable, configure, and verify VRRP in IBM Networking OS switches, complete the following steps:

1. Enable VRRP.

   Run **router vrrp** to enter the VRRP configuration mode and **enable** to activate the protocol on both AGG-1 and AGG-2 switches (Example 6-29).

   Example 6-29 Enable VRRP

```
AGG-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-1(config)#router vrrp
AGG-1(config-vrrp)#enable

AGG-2#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-2(config)#router vrrp
AGG-2(config-vrrp)#enable
```

2. Configure the virtual router:

   a. Configure the virtual router ID (Example 6-30).

      Up to 15 virtual router instances can be defined in IBM Networking OS V6.8.

      Run **virtual-router <1-15> virtual-router-id <1-255>** in VRRP configuration mode to define the the virtual router ID (VRID). To create a pool of VRRP-enabled routing devices that can provide redundancy to each other, each participating VRRP device must be configured with the same virtual router.

      The VRID for standard virtual routers (where the virtual router IP address is not the same as any virtual server) can be any integer 1 - 255. The default value is 1. All VRID values must be unique within the VLAN to which the virtual router's IP interface belongs.

      Example 6-30 Virtual router ID configuration

```
AGG-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-1(config)#router vrrp
AGG-1(config-vrrp)#virtual-router 3 virtual-router-id 30
AGG-1(config-vrrp)#

AGG-2#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-2(config)#router vrrp
AGG-2(config-vrrp)#virtual-router 3 virtual-router-id 30
AGG-2(config-vrrp)#
```

   b. Configure the virtual router IP address (Example 6-31).

      Run **[no] virtual-router <1-15> address <IP address>** to define an IP address for this virtual router by using dotted decimal notation. This address is used in conjunction with the VRID to configure the same virtual router on each participating VRRP device. The default address is 0.0.0.0.

      Example 6-31 Virtual router IP address configuration

```
AGG-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-1(config)#router vrrp
```

```
AGG-1(config-vrrp)#virtual-router 3 address 10.0.30.1
AGG-1(config-vrrp)#

AGG-2#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-2(config)#router vrrp
AGG-2(config-vrrp)#virtual-router 3 address 10.0.30.1
AGG-2(config-vrrp)#
```

c. Select a switch IP interface (Example 6-32).

Run **virtual-router <1-15> interface <interface number>** to select a switch IP interface. If the IP interface has the same IP address as the **addr** option, this switch is considered the "owner" of the defined virtual router. An owner has a special priority of 255 (highest) and always assumes the role of Master router, even if it must pre-empt another virtual router that assumed Master routing authority. This preemption occurs even if the **preem** option is disabled. The default value is 1.

Example 6-32 IP interface selection

```
AGG-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-1(config)#router vrrp
AGG-1(config-vrrp)#virtual-router 3 interface 30
AGG-1(config-vrrp)#

AGG-2#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-2(config)#router vrrp
AGG-2(config-vrrp)#virtual-router 3 interface 30
AGG-2(config-vrrp)#
```

d. Define the election priority (Example 6-33).

Run **virtual-router <1-15> priority <1-254>** to define the election priority bias for this virtual server. The priority value can be any integer 1 - 254. The default value is 100.

During the Master router election process, the routing device with the highest virtual router priority number wins. If there is a tie, the device with the highest IP interface address wins. If this virtual router's IP address is the same as the one used by the IP interface, the priority for this virtual router is automatically set to 255 (highest).

When priority tracking is used, this base priority value can be modified according to a number of performance and operational criteria.

In our reference architecture, we used a virtual router IP address that is different from the IP interface addresses and assigned a higher priority to the AGG-1 switch, in order for AGG-1 to become an elected Master. The AGG-2 switch was left at its default priority (100) (Example 6-33).

Example 6-33 Virtual router election priority configuration

```
AGG-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-1(config)#router vrrp
AGG-1(config-vrrp)#virtual-router 3 priority 105
AGG-1(config-vrrp)#
```

e. Configure preemption (Example 6-34).

Run **[no] virtual-router <1-15> preemption** to enable or disable master preemption. When enabled, if this virtual router is in backup mode but has a higher priority than the current Master, this virtual router pre-empts the lower priority Master and assumes control. Even when preemption is disabled, this virtual router always preempts any other Master if this switch is the owner (the IP interface address and virtual router **addr** are the same). By default, this option is enabled.

Example 6-34 Configure preemption

```
AGG-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-1(config)#router vrrp
AGG-1(config-vrrp)#virtual-router 3 preemption
AGG-1(config-vrrp)#

AGG-2#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-2(config)#router vrrp
AGG-2(config-vrrp)#virtual-router 3 preemption
AGG-2(config-vrrp)#
```

f. Configure the timers, as shown in Example 6-35.

Run **virtual-router <1-15> timers advertise <1-255>** to define the time interval between VRRP master advertisements. This value can be any integer 1 - 255 seconds. The default value is 1.

Run **virtual-router <1-15> timers preempt-delay-time <0-255>** to configure the preempt delay interval. This timer is configured on the VRRP Owner and prevents the switch from moving back to the Master state until the preempt delay interval expires. Ensure that the interval is long enough for OSPF or other routing protocols to converge.

Run **[no] virtual-router <1-128> fast-advertise** to enable or disable Fast Advertisements. When enabled, the VRRP master advertisements interval is calculated in units of centiseconds, instead of seconds. For example, if **adver** is set to 1 and Fast Advertisement is enabled, master advertisements are sent every 0.01 second. When you disable fast advertisement, the advertisement interval is set to the default value of 1 second. To support Fast Advertisements, set the interval to 20 - 100 centiseconds.

Important: Fast Advertisements were not used in reference architecture implementation.

Example 6-35 Virtual router timers configuration

```
AGG-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-1(config)#router vrrp
AGG-1(config-vrrp)#virtual-router 3 timers advertise 2
AGG-1(config-vrrp)#virtual-router 3 timers preempt-delay-time 5
AGG-1(config-vrrp)#

AGG-2#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-2(config)#router vrrp
```

```
AGG-2(config-vrrp)#virtual-router 3 timers advertise 2
AGG-2(config-vrrp)#virtual-router 3 timers preempt-delay-time 5
AGG-2(config-vrrp)#
```

g. Enable the configured virtual router.

Run **[no] virtual-router <1-15> enable** to enable or disable this virtual router.

Run **no virtual-router <1-15>** to delete this virtual router from the switch configuration.

Note: We enabled the AGG-2 (Backup) router first to show the preemption operation. In Example 6-36, note the log messages that show the role status change. Observe the time stamps of the messages. You can see that AGG-1 assumed the Master role and AGG-2 returned to Backup state after the virtual router is enabled on AGG-1 with the preemption option enabled.

Example 6-36 Virtual router activation with preemption enabled

```
AGG-2#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-2(config)#router vrrp
AGG-2(config-vrrp)#virtual-router 3 enable
AGG-2(config-vrrp)#

Aug 16 11:26:57 AGG-2 NOTICE  vrrp: virtual router       10.0.30.1 is now BACKUP
Aug 16 11:27:04 AGG-2 NOTICE  vrrp: virtual router       10.0.30.1 is now MASTER.

Aug 16 23:45:22 AGG-2 NOTICE  vrrp: virtual router       10.0.30.1 is now BACKUP

AGG-2(config-vrrp)#^Z
AGG-2#
AGG-2#


AGG-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-1(config)#router vrrp
AGG-1(config-vrrp)#virtual-router 3 enable
AGG-1(config-vrrp)#

Aug 16 11:27:19 AGG-1 NOTICE  vrrp: virtual router       10.0.30.1 is now BACKUP
Aug 16 11:27:24 AGG-1 NOTICE  vrrp: virtual router       10.0.30.1 is now MASTER.

AGG-1(config-vrrp)#^Z

AGG-1#
```

3. Configure tracking.

The commands shown in this step are used to modify the priority system when electing the Master router from a pool of virtual routers. Various tracking criteria can be used to bias the election results. Each time one of the tracking criteria is met, the priority level for the virtual router is increased by an amount defined through the VRRP Tracking commands. Criteria are tracked dynamically, continuously updating virtual router priority levels when enabled. If the virtual router preemption option is enabled, this virtual router can assume Master routing authority when its priority level rises above the priority level of the current Master.

Some tracking criteria apply to standard virtual routers called *virtual interface routers*. A virtual server router is defined as any virtual router whose IP address is the same as any configured virtual server IP address.

Run **[no] virtual-router <1-15> track virtual-routers** to allow the priority for this virtual router to be increased for each virtual router in Master mode on this switch. This command is useful for making sure that traffic for any particular client/server pairing is handled by the same switch, increasing routing and load balancing efficiency. This command is disabled by default.

Run **[no] virtual-router <1-15> track interfaces** to allow the priority for this virtual router to be increased for each other IP interface active on this switch. An IP interface is considered active when there is at least one active port on the same VLAN. This command helps elect the virtual routers with the most available routes as the Master. This command is disabled by default.

Run **[no] virtual-router <1-15> track ports** to allow the priority for this virtual router to be increased for each active port on the same VLAN. A port is considered “active” if it has a link and is forwarding traffic. This command helps elect the virtual routers with the most available ports as the Master. This command is disabled by default.

Use the following commands to set weights for the various criteria used to modify priority levels during the Master router election process. Each time one of the tracking criteria is met, the priority level for the virtual router is increased by a defined amount. These priority tracking options define only increment values. These options do not affect the VRRP Master router election process until options under the VRRP Virtual Router Priority Tracking Commands are enabled.

Run **tracking-priority-increment virtual-routers <0-254>** to define the priority increment value (0 - 254) for virtual routers in Master mode detected on this switch. The default value is 2.

Run **tracking-priority-increment interfaces <0-254>** to define the priority increment value for active IP interfaces detected on this switch. The default value is 2.

Run **tracking-priority-increment ports <0-254>** to define the priority increment value for active ports on the virtual router’s VLAN. The default value is 2.

The tracking configuration is shown in Example 6-37.

Example 6-37 Configure tracking

```
AGG-1#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-1(config)#router vrrp
AGG-1(config-vrrp)#virtual-router 3 track ports
AGG-1(config-vrrp)#tracking-priority-increment ports 50
AGG-1(config-vrrp)#

AGG-2#configure terminal
Enter configuration commands, one per line.  End with Ctrl/Z.
AGG-2(config)#router vrrp
AGG-2(config-vrrp)#virtual-router 3 track ports
AGG-2(config-vrrp)#tracking-priority-increment ports 50
AGG-2(config-vrrp)#
```

Important: For other VRRP configuration commands and details, see 6.7, “More information” on page 284.

4. Verify the VRRP operation

   Run the commands listed in this step to verify the VRRP operation on the switch.

   Run **show ip vrrp** to display the current VRRP parameters (Example 6-38).

Example 6-38 Verify the VRRP current parameters

```
AGG-1#show ip vrrp
Current VRRP settings: ON
Current VRRP hold off time: 0

Current VRRP Tracking settings:
   vrs 2, ifs 2, ports 50

Current VRRP Virtual Router Group:
    vrid 1, if 1, prio 100, adver 1, disabled
    preem enabled, fast-advertisement disabled
    track nothing

Current VRRP virtual router settings:
  3: vrid 30, 10.0.30.1, if 30, prio 105, adver 2, enabled
    preem enabled, predelay 5, fast-advertisement disabled
    track ports
AGG-1#

AGG-2#show ip vrrp
Current VRRP settings: ON
Current VRRP hold off time: 0

Current VRRP Tracking settings:
   vrs 2, ifs 2, ports 50

Current VRRP Virtual Router Group:
    vrid 1, if 1, prio 100, adver 1, disabled
    preem enabled, fast-advertisement disabled
    track nothing

Current VRRP virtual router settings:
  3: vrid 30, 10.0.30.1, if 30, prio 100, adver 2, enabled
    preem enabled, predelay 5, fast-advertisement disabled
    track ports
AGG-2#
```

Run **show ip vrrp virtual-router <1-15>** to display the current VRRP parameters of the selected virtual router (Example 6-39).

Example 6-39 Verify the selected virtual router current parameters

```
AGG-1#show ip vrrp virtual-router 3
Current VRRP virtual router 3:
 vrid 30, 10.0.30.1, if 30, prio 105, adver 2, enabled
    preem enabled, predelay 5, fast-advertisement disabled
    track ports
AGG-1#


AGG-2#show ip vrrp virtual-router 3
```

```
Current VRRP virtual router 3:
 vrid 30, 10.0.30.1, if 30, prio 100, adver 2, enabled
    preem enabled, predelay 5, fast-advertisement disabled
    track ports
AGG-2#
```

Run **show ip vrrp counters** to display VRRP statistics (Example 6-40).

Example 6-40 Show VRRP statistics

```
AGG-2#show ip vrrp counters
VRRP statistics:
vrrpInAdvers:          79786   vrrpBadAdvers:                 25
vrrpOutAdvers:           174   vrrpOutGratuitousARPs:          2
vrrpBadVersion:            0   vrrpBadVrid:                    0
vrrpBadAddress:            0   vrrpBadData:                    0
vrrpBadPassword:           0   vrrpBadInterval:               25
AGG-2#
```

Run **show ip vrrp track-priority-increment** to display VRRP tracking priority increments configuration (Example 6-41).

Example 6-41 Show VRRP tracking priority increments

```
AGG-1#show ip vrrp track-priority-increment
Current VRRP Tracking settings:
   vrs 2, ifs 2, ports 50
AGG-1#
```

Run **show ip vrrp virtual-router <1-15> track** to display the selected virtual router tracking configuration.

Example 6-42 Show the virtual router tracking configuration

```
AGG-1#show ip vrrp virtual-router 3 track
Current VRRP virtual router 3 tracking:
       track ports
AGG-1#
```

6.6 IPv4 and IPv6 end-to-end connectivity verification

With both Top-of-Rack and embedded switches architecture implemented, our test host, connected at the ends of the network, is now reachable and able to communicate on IPv4 and IPv6.

Figure 6-3 shows the complete architecture with the IP addressing details. SRV-1 is connected to ACC-1 and ACC-2 and is able to **`ping`** the SRV-3 connected to the ACC-3 stack.

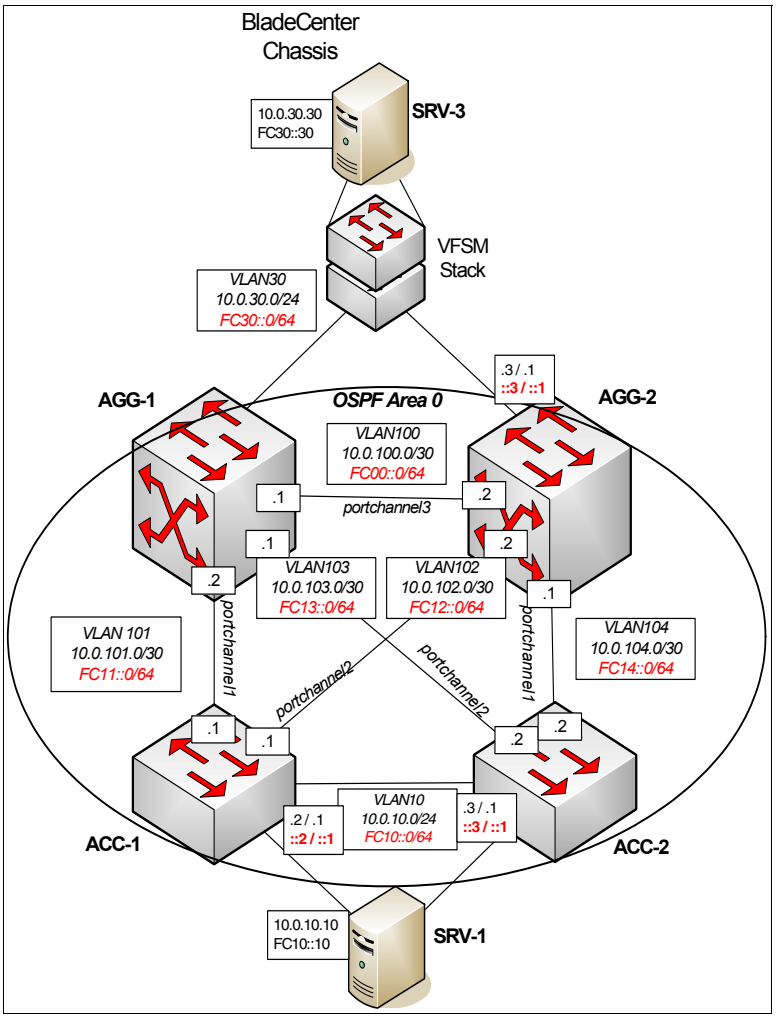

Figure 6-3 Complete network topology

The details of SRV-1 and SRV-3 are as follows:

- SRV-1: 10.0.10.10 / FC10::10 (Windows Server 2008 R2)
- SRV-3: 10.0.30.30 / FC30::30 (Red Hat Enterprise Linux)

Windows host verification

Figure 6-4 shows that the Windows host is able to **ping** the Linux host using both IPv4 and IPv6.

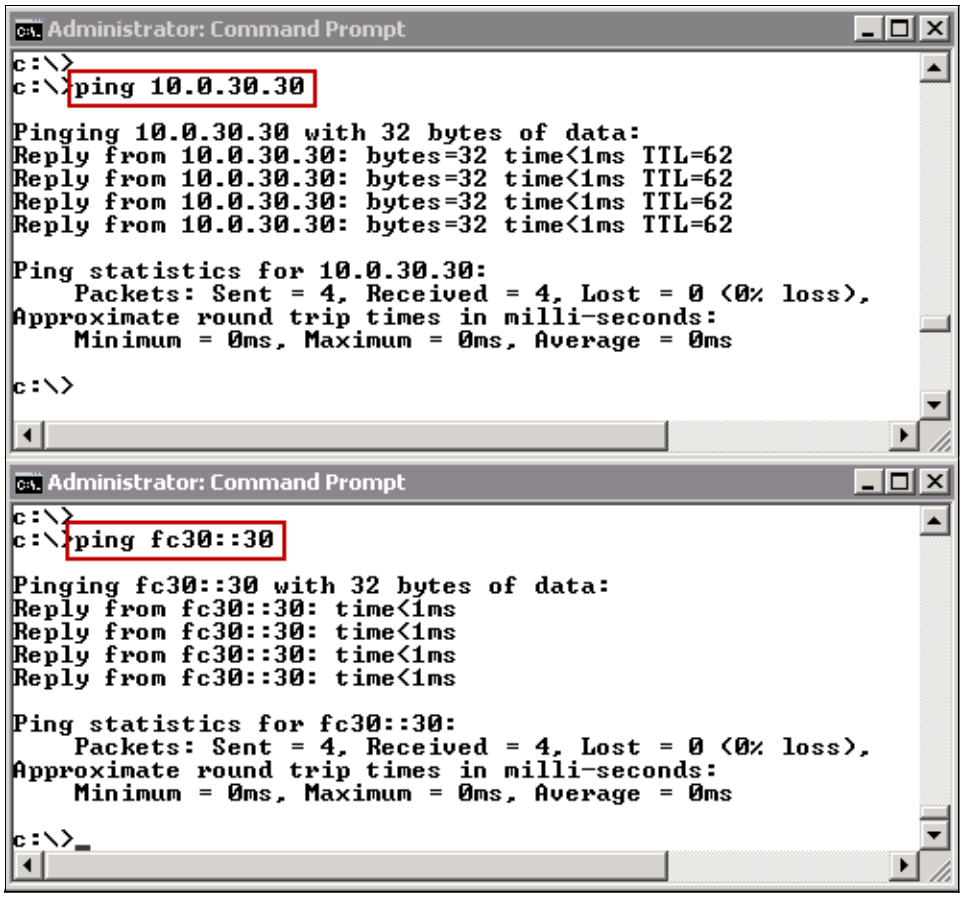

Figure 6-4 Windows host to Linux host verification

Linux host verification

Figure 6-5 shows the output from an ICMP test from a Linux host to a Windows host.

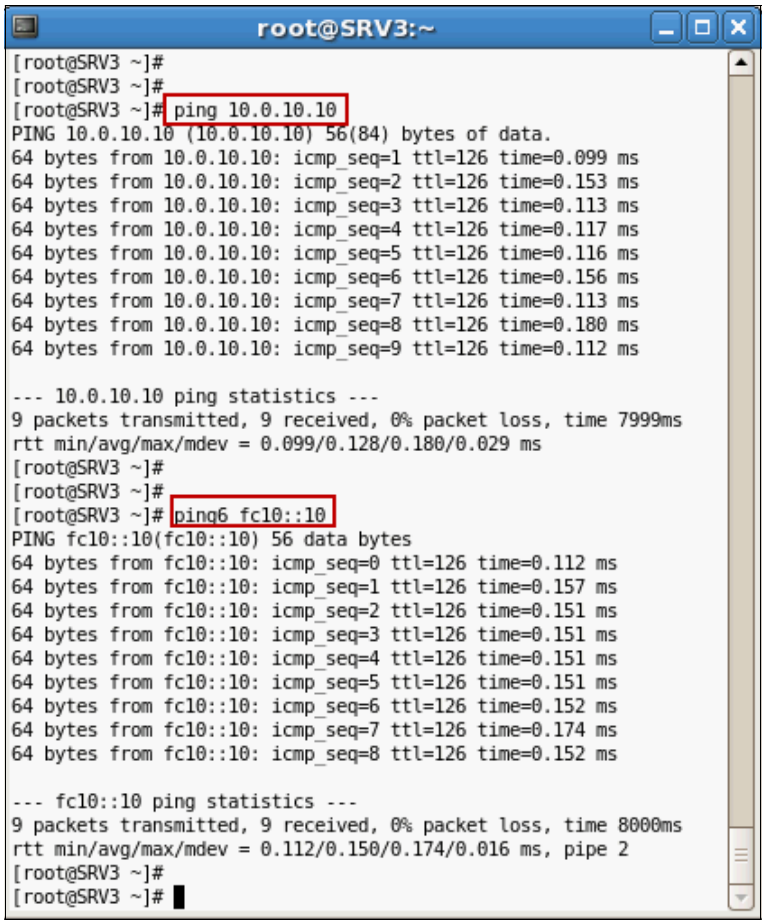
root@SRV3:~
[root@SRV3 ~]#
[root@SRV3 ~]#
[root@SRV3 ~]# ping 10.0.10.10
PING 10.0.10.10 (10.0.10.10) 56(84) bytes of data.
64 bytes from 10.0.10.10: icmp_seq=1 ttl=126 time=0.099 ms
64 bytes from 10.0.10.10: icmp_seq=2 ttl=126 time=0.153 ms
64 bytes from 10.0.10.10: icmp_seq=3 ttl=126 time=0.113 ms
64 bytes from 10.0.10.10: icmp_seq=4 ttl=126 time=0.117 ms
64 bytes from 10.0.10.10: icmp_seq=5 ttl=126 time=0.116 ms
64 bytes from 10.0.10.10: icmp_seq=6 ttl=126 time=0.156 ms
64 bytes from 10.0.10.10: icmp_seq=7 ttl=126 time=0.113 ms
64 bytes from 10.0.10.10: icmp_seq=8 ttl=126 time=0.180 ms
64 bytes from 10.0.10.10: icmp_seq=9 ttl=126 time=0.112 ms
--- 10.0.10.10 ping statistics ---
9 packets transmitted, 9 received, 0% packet loss, time 7999ms
rtt min/avg/max/mdev = 0.099/0.128/0.180/0.029 ms
[root@SRV3 ~]#
[root@SRV3 ~]#
[root@SRV3 ~]# ping6 fc10::10
PING fc10::10(fc10::10) 56 data bytes
64 bytes from fc10::10: icmp_seq=0 ttl=126 time=0.112 ms
64 bytes from fc10::10: icmp_seq=1 ttl=126 time=0.157 ms
64 bytes from fc10::10: icmp_seq=2 ttl=126 time=0.151 ms
64 bytes from fc10::10: icmp_seq=3 ttl=126 time=0.151 ms
64 bytes from fc10::10: icmp_seq=4 ttl=126 time=0.151 ms
64 bytes from fc10::10: icmp_seq=5 ttl=126 time=0.151 ms
64 bytes from fc10::10: icmp_seq=6 ttl=126 time=0.152 ms
64 bytes from fc10::10: icmp_seq=7 ttl=126 time=0.174 ms
64 bytes from fc10::10: icmp_seq=8 ttl=126 time=0.152 ms
--- fc10::10 ping statistics ---
9 packets transmitted, 9 received, 0% packet loss, time 8000ms
rtt min/avg/max/mdev = 0.112/0.150/0.174/0.016 ms, pipe 2
[root@SRV3 ~]#
[root@SRV3 ~]#

Figure 6-5 Linux host to Windows host verification

6.7 More information

For detailed information about Layer 2 configuration, see the following documents:

- Configuration guides:

  IBM Virtual Fabric 10Gb Switch Module for IBM BladeCenter Application Guide (6.8):

  http://www-01.ibm.com/support/docview.wss?uid=isg3T7000461

- Command reference guides:

  IBM Virtual Fabric 10G Switch Module Industry-Standard CLI Reference (6.8):

  http://www-01.ibm.com/support/docview.wss?uid=isg3New

7

Maintenance and troubleshooting

In this chapter, we describe some elements that can help you with the maintenance and troubleshooting of IBM System Networking 10Gb switches.

© Copyright IBM Corp. 2012. All rights reserved.

7.1 Configuration management

This section describes how to manage configuration files, save and restore a configuration in the switch, perform a firmware upgrade, and identify some problems by checking the system logs and descriptions.

7.1.1 Configuration files

The switch stores its configuration in two files:

- `startup-config` is the configuration the switch uses when it is reloaded.
- `running-config` is the configuration that reflects all the changes you made from the CLI. It is stored in memory and is lost after the reload of the switch.

7.1.2 Configuration blocks

The switch stores its configuration in one of two configuration blocks, or partitions:

- `active-config` is the active configuration file.
- `backup-config` is the alternative configuration file.

This setup has the flexibility you need to manage the configuration of the switch and perform a possible configuration rollback.

7.1.3 Managing configuration files

This section describes the different ways of managing the configuration files.

Managing the configuration using the CLI

You can manage the configuration files by using several commands:

- Run the following command to dump the current configuration file:

  ```
  Switch#show running-config
  ```

- Run the following command to copy the current (running) configuration from switch memory to the `backup-config` partition:

  ```
  Switch#copy running-config backup-config
  ```

- Run the following command to copy the current (running) configuration from switch memory to the `startup-config` partition:

  ```
  Switch#copy running-config startup-config
  ```

- Run the following command to copy the current (running) configuration from switch memory to the `active-config` partition:

  ```
  Switch#write memory
  ```

- Run the following command to back up the current configuration to a file on the selected FTP/TFTP server:

  ```
  Switch#copy running-config {ftp|tftp}
  ```

- Run the following command to restore the current configuration from an FTP/TFTP server.

  ```
  Switch#copy {ftp|tftp} running-config
  ```

Managing the configuration through SNMP

This section describes how to use MIB calls to work with switch images and configuration files.

You can use a standard SNMP tool to perform the actions, using the MIBs listed in Table 7-1. For information about how to set up your switch to use SNMP, see 7.3.2, “SNMP” on page 297.

Table 7-1 SNMP MIBs for managing switch configuration and firmware

MIB name	MIB OID
agTransferServer	1.3.6.1.4.1872.2.5.1.1.7.1.0
agTransferImage	1.3.6.1.4.1872.2.5.1.1.7.2.0
agTransferImageFileName	1.3.6.1.4.1872.2.5.1.1.7.3.0
agTransferCfgFileName	1.3.6.1.4.1872.2.5.1.1.7.4.0
agTransferDumpFileName	1.3.6.1.4.1872.2.5.1.1.7.5.0
agTransferAction	1.3.6.1.4.1872.2.5.1.1.7.6.0
agTransferLastActionStatus	1.3.6.1.4.1872.2.5.1.1.7.7.0
agTransferUserName	1.3.6.1.4.1872.2.5.1.1.7.9.0
agTransferPassword	1.3.6.1.4.1.1872.2.5.1.1.7.10.0
agTransferTSDumpFileName	1.3.6.1.4.1.1872.2.5.1.1.7.11.0

The following SNMP actions can be performed by using the MIBs listed in Table 7-1:

- Load a new Switch image (boot or running) from an FTP/TFTP server.
- Load a previously saved switch configuration from an FTP/TFTP server.
- Save the switch configuration to an FTP/TFTP server.
- Save a switch memory dump to an FTP/TFTP server.

Loading a new switch image

To load a new switch image with the name `MyNewImage-1.img` into image2, complete the following steps. This example shows an FTP/TFTP server at IPv4 address `172.25.101.200`, although IPv6 is also supported.

1. Set the FTP/TFTP server address where the configuration file is saved:

   ```
   Set agTransferServer.0 "172.25.101.200".
   ```

2. Set the name of the configuration file:

   ```
   Set agTransferCfgFileName.0 "MyRunningConfig.cfg"
   ```

3. If you are using an FTP server, enter a user name:

   ```
   Set agTransferUserName.0 "MyName"
   ```

4. If you are using an FTP server, enter a password:

   ```
   Set agTransferPassword.0 "MyPassword"
   ```

5. Initiate the transfer. To save a running configuration file, enter 4.

   ```
   Set agTransferAction.0 "4"
   ```

Loading a saved configuration

To load a saved switch configuration with the name `MyRunningConfig.cfg` into the switch, complete the following steps. This example shows a TFTP server at IPv4 address `172.25.101.200`, although IPv6 is also supported, where the configuration previously saved is available to download.

1. Set the FTP/TFTP server address where the switch Configuration File is located:

   `Set agTransferServer.0 "172.25.101.200"`

2. Set the name of the configuration file:

   `Set agTransferCfgFileName.0 "MyRunningConfig.cfg"`

3. If you are using an FTP server, enter a user name:

   `Set agTransferUserName.0 "MyName"`

4. If you are using an FTP server, enter a password:

   `Set agTransferPassword.0 "MyPassword"`

5. Initiate the transfer. To restore a running configuration, enter 3.

   `Set agTransferAction.0 "3"`

Saving the configuration

To save the switch configuration to an FTP/TFTP server, complete the following steps. This example shows an FTP/TFTP server at IPv4 address `172.25.101.200`, although IPv6 is also supported.

1. Set the FTP/TFTP server address where the configuration file is saved:

   `Set agTransferServer.0 "172.25.101.200"`

2. Set the name of the configuration file:

   `Set agTransferCfgFileName.0 "MyRunningConfig.cfg"`

3. If you are using an FTP server, enter a user name:

   `Set agTransferUserName.0 "MyName"`

4. If you are using an FTP server, enter a password:

   `Set agTransferPassword.0 "MyPassword"`

5. Initiate the transfer. To save a running configuration file, enter 4.

   `Set agTransferAction.0 "4"`

Saving a switch memory dump

You can also save a switch memory dump, which has information about the switch. To save a switch memory dump to an FTP/TFTP server, complete the following steps. This example shows an FTP/TFTP server at `172.25.101.200`, although IPv6 is also supported.

1. Set the FTP/TFTP server address where the configuration is saved:

   `Set agTransferServer.0 "172.25.101.200"`

2. Set the name of the dump file:

   `Set agTransferDumpFileName.0 "MyDumpFile.dmp"`

3. If you are using an FTP server, enter a user name:

   `Set agTransferUserName.0 "MyName"`

4. If you are using an FTP server, enter a password:

   ```
   Set agTransferPassword.0 "MyPassword"
   ```

5. Initiate the transfer. To save a dump file, enter 5.

   ```
   Set agTransferAction.0 "5"
   ```

7.1.4 Factory defaults

To reset the switch to the factory defaults, you need to perform one of the following procedures.

Resetting with access to the terminal

If you have access to the switch's terminal, and you would like to reset the switch to the factory defaults, complete the following steps:

1. Set the configuration block to "factory":

   ```
   RS8264(config)#boot configuration-block factory
   .
   .
   Next boot will use factory default config block instead of active.
   ```

2. Reload the switch:

   ```
   RSG8264#reload

   Reset will use software "image2" and the active config block.
   >> Note that this will RESTART the Spanning Tree,
   >> which will likely cause an interruption in network service.
   Confirm reload (y/n) ? y
   ```

Resetting with no access to the terminal

If you want to reset the switch to the factory defaults and have no access to the terminal, complete the following steps:

1. You need to reload the switch by powering it off or by running **reload**.
2. You can interrupt the boot process and enter the Boot Management menu from the serial console port. When the system shows Memory Test, press Shift+B. The Boot Management menu opens:

   ```
   Resetting the System ...
   Memory Test ..............................
   Boot Management Menu
   1 - Change booting image
   2 - Change configuration block
   3 - Xmodem download
   4 - Exit
   Please choose your menu option:
   ```

3. Enter 2 to change the configuration block:

   ```
   Please choose your menu option: 2

   Unknown current config block 255
   Enter configuration block: a, b or f (active, backup or factory):
   ```

4. Enter f to use the factory defaults:

   ```
   Enter configuration block: a, b or f (active, backup or factory): f
   ```

5. You see the initial menu once again. Enter `4` to exit and reset the switch with the default configuration:

```
Boot Management Menu
        1 - Change booting image
        2 - Change configuration block
        3 - Xmodem download
        4 - Exit
Please choose your menu option:4
```

The switch resets to the factory default configuration.

7.1.5 Password recovery

To perform password recovery, you need to set the switch to the factory default by using one of the procedures described in 7.1.4, “Factory defaults” on page 289.

After you reset the switch, run the following command:

```
RS8264#copy active-config running-config
```

After the command finishes running, the switch is in enable mode without a password. Change the password by running **`password`** in configuration mode:

```
RS8264(config)#password
```

Important: If you set the configuration block to `factory` from CLI, do not forget to change it back to either active or backup configuration by running the following command:

```
RS8264(config)#boot configuration-block active
```

7.2 Firmware management

The switch software image is the executable code that runs on the switch. A version of the image comes preinstalled on the device. As new versions of the image are released, you can upgrade the software that runs on your switch. To obtain the latest version of software supported for your switch, go to the IBM Support website, found at:

http://www.ibm.com/support

7.2.1 Firmware files

IBM switches can store up to two different versions of the switch software, or OS, images (called `image1` and `image2`) and special boot software, or a boot image (called `boot`). When you load new software, make sure that you upgrade both OS and boot images.

In IBM Networking OS, run **`show boot`** to see what images are installed. The output is shown in Example 7-1.

Example 7-1 Showing the current version of the boot image on the switch

```
RS8264#show boot
Currently set to boot software image1, active config block.
NetConfig: disabled, NetConfig tftp server:  , NetConfig cfgfile:
Currently set to boot with default profile
Current CLI mode set to ISCLI with selectable prompt disabled.
```

```
Current FLASH software:
  image1: version 6.3.2, downloaded  7:36:34 Tue Jan  3, 2000
  image2: version 6.8.0.3, downloaded 11:38:34 Fri Jan 20, 2000
  boot kernel: version 6.8.0.3
Currently scheduled reboot time: none
```

In Example 7-1 on page 290, you can see that the system has two OS images:

- `image1`: Version 6.3.2
- `image2`: Version 6.8.0.3

The boot image version is 6.8.0.3. We want to make sure that the switch uses the same version for boot image and OS image. To change the boot image, run the following command:

```
RS8264(config)#boot image image2
Next boot will use switch software image2 instead of image1.
```

7.2.2 Boot Management Menu

Before getting into the details about how the firmware is loaded, we describe the Boot Management Menu. You can use the Boot Management Menu to switch the software image, reset the switch to factory defaults, or recover from a failed software download.

You can interrupt the boot process and enter the Boot Management Menu from the serial console port. When the system shows Memory Test, press Shift + B. The Boot Management Menu then appears.

The actual menu shown is similar to Example 7-2, in which we show what happens when we decide to change the boot image by pressing `1` and then choosing a secondary image by pressing `2`.

Example 7-2 Boot Management Menu

```
Resetting the System ...
Memory Test ..............................
Boot Management Menu
1 - Change booting image
2 - Change configuration block
3 - Xmodem download
4 - Exit
Please choose your menu option: 1
Current boot image is 1. Enter image to boot: 1 or 2: 2
Booting from image 2
```

We refer to this menu while describing how to load a firmware, or restore or recover from a failed installation.

Important: When connecting to the switch with the serial port, you should use the following parameters, which do not change, except for the speed, that changes when we recover from a faulty upgrade:

- Speed: 9600 bps
- Data Bits: 8
- Stop Bits: 1
- Parity: None
- Flow Control: None

7.2.3 Loading the new firmware

In this section, we show how to load the new firmware on the switch by using both Menu-Based CLI and ISCLI.

The typical upgrade process for the upgrade of software image has the following steps:

1. Load a new software image and boot image onto an FTP or TFTP server on your network.
2. Transfer the new images to your switch.
3. Specify the new software image as the one that is loaded into switch memory the next time a switch reset occurs.
4. Reset the switch.

Loading the new firmware with the Menu-Based CLI

To load the new firmware with the Menu-Based CLI, complete the following steps:

1. Enter the following Boot Options command from the CLI:

   ```
   >> # /boot/gtimg
   ```

2. Enter the name of the switch software to be replaced when prompted:

   ```
   Enter name of switch software image to be replaced
   ["image1"/"image2"/"boot"]: <image>
   ```

3. Enter the host name or IP address of the FTP or TFTP server:

   ```
   Enter hostname or IP address of FTP/TFTP server: <hostname or IP address>
   ```

4. Enter the name of the new software file on the server:

   ```
   Enter name of file on FTP/TFTP server: <filename>
   ```

   The exact form of the name varies by server. However, the file location is normally relative to the FTP or TFTP directory (`/tftpboot`).

5. Enter your user name for the server, if applicable:

   ```
   Enter username for FTP server or hit return for
   TFTP server: {<username>|<Enter>}
   ```

   If entering an FTP server user name, you are also prompted for the password. The system then prompts you to confirm your request. Once confirmed, the software loads into the switch.

6. If software is loaded into a different image than the one most recently booted, the system prompts you whether you want to run the new image at next boot. Otherwise, you can enter the **`image`** command at the `Boot Options#` prompt:

   ```
   Boot Options# image
   ```

7. The system then informs you which software image (`image1` or `image2`) is currently set to be loaded at the next reset, and prompts you to enter a new choice:

   ```
   Currently set to use switch software "image1" on next reset.
   Specify new image to use on next reset ["image1"/"image2"]:
   ```

   Specify the image that contains the newly loaded software.

8. Reboot the switch to run the new software by running **`reset`**:

   ```
   Boot Options# reset
   ```

   The system prompts you to confirm your request. Once confirmed, the switch reboots to use the new software.

Loading the firmware with ISCLI

To load the firmware with ISCLI, complete the following steps:

1. To determine the software version currently used on the switch, run **`show boot`**:

   ```
   RS8264# show boot
   ```

2. In Privileged EXEC mode, run **`copy`**:

   ```
   Router# copy {tftp|ftp} {image1|image2|boot-image}
   ```

3. Enter the host name or IP address of the FTP or TFTP server:

   ```
   Address or name of remote host: <name or IP address>
   ```

4. Enter the name of the new software file on the server:

   ```
   Source file name: <filename>
   ```

   The exact form of the name varies by server. However, the file location is normally relative to the FTP or TFTP directory (for example, `tftpboot`).

5. If required by the FTP or TFTP server, enter the appropriate user name and password. The switch prompts you to confirm your request. Once confirmed, the software begins loading into the switch.

6. When loading is complete, run the following commands to enter Global Configuration mode to select which software image (image1 or image2) you want to run in switch memory for the next reboot:

   ```
   Router# configure terminal
   Router(config)# boot image {image1|image2}
   ```

7. The system then verifies which image is set to be loaded at the next reset:

   ```
   Next boot will use switch software image1 instead of image2.
   ```

8. Reboot the switch to run the new software by running **`reload`**:

   ```
   Router(config)# reload
   ```

The system prompts you to confirm your request. Once confirmed, the switch reboots to use the new software version installed.

7.2.4 Recovering from a failed firmware upgrade

Rarely, the firmware upgrade process fails. This situation is unlikely to happen, but if it occurs, you can still recover from this situation. To recover your switch, connect a PC to the serial port of your switch while the switch is off, and access the switch as described in the appropriate User Guide (see “Related publications” on page 333).

Important: The procedure described in this section might also be useful when you boot the switch and the boot and OS versions are not equal.

Then, power on the switch and you see some boot messages. From your terminal window, press Shift + B while the Memory tests are processing and dots are showing the progress.

A menu opens. Select `3` for `Xmodem download`. Change the serial connection properties as follows:

```
## Switch baudrate to 115200 bps and press ENTER ...
```

Change the settings of your terminal to meet the 115300 bps requirement and press Enter. The system switches to download accept mode. You see a series of “C” characters in the screen that prompt you when the switch is ready. Start an Xmodem terminal to push the boot code you want to restore into the switch. Select the boot code for your system, and the switch start the download. You should see a screen similar to Example 7-3.

Example 7-3 Xmodem firmware download

```
yzModem - CRC mode, 62494(SOH)/0(STX)/0(CAN) packets, 6 retries
Extracting images ... Do *NOT* power cycle the switch.
**** VMLINUX ****
Un-Protected 10 sectors
Erasing Flash.............. done
Writing to Flash.............done
Protected 10 sectors
**** RAMDISK ****
Un-Protected 44 sectors
Erasing Flash.............................................. done
Writing to Flash..............................................done
Protected 44 sectors
**** BOOT CODE ****
Un-Protected 8 sectors
Erasing Flash........... done
Writing to Flash...........done
Protected 8 sectors
```

When this process is finished, you are prompted to reconfigure your terminal to 9600 bps speed:

```
## Switch baudrate to 9600 bps and press ESC ...
```

Change the speed of your serial connection, and then press Esc. The Boot Management Menu opens again. Select `3` again, and change the speed to 115000 bps when the following message appears, to start pushing the OS image.

```
## Switch baudrate to 115200 bps and press ENTER ...
```

After you are done, press Enter to continue download. Select the OS image you want to upload to the switch. The Xmodem client starts sending the image to the switch. When the upload is complete, you see a screen similar to the one in Example 7-4.

Example 7-4 OS image upgrade

```
yzModem - CRC mode, 27186(SOH)/0(STX)/0(CAN) packets, 6 retries
Extracting images ... Do *NOT* power cycle the switch.
```

```
**** Switch OS ****
Please choose the Switch OS Image to upgrade [1|2|n] :
```

You are prompted to the select the image space in the switch you want to upgrade. If you are performing a recovery, select `1`. You see a screen similar to the one in Example 7-5.

Example 7-5 Upgrading the OS image

```
Switch OS Image 1 ...
Un-Protected 27 sectors
Erasing Flash.............................. done
Writing to Flash.............................done
Protected 27 sectors
```

When this process is done, You are prompted to reconfigure your terminal to 9600 bps speed again:

```
## Switch baudrate to 9600 bps and press ESC ...
```

Press Esc to show the Boot Management Menu, and choose option `4` to exit and boot the new image.

7.3 Logging and reporting

This section focuses on how to manage and configure systems logs, and how to configure an SNMP agent and SNMP traps.

7.3.1 System logs

IBM Networking OS can provide some information through a system log that uses the following syntax when outputting system log (syslog) messages:

```
<Time stamp><Log Label>BLADEOS<Thread ID>:<Message>
```

You can view the latest system logs by running **`show logging messages`** (Example 7-6).

Example 7-6 Example of syslog output

```
Jul 8 17:25:41 NOTICE system: link up on port 1
Jul 8 17:25:41 NOTICE system: link up on port 8
Jul 8 17:25:41 NOTICE system: link up on port 7
Jul 8 17:25:41 NOTICE system: link up on port 2
Jul 8 17:25:41 NOTICE system: link up on port 1
Jul 8 17:25:41 NOTICE system: link up on port 4
Jul 8 17:25:41 NOTICE system: link up on port 3
Jul 8 17:25:41 NOTICE system: link up on port 6
Jul 8 17:25:41 NOTICE system: link up on port 5
Jul 8 17:25:41 NOTICE system: link up on port 4
Jul 8 17:25:41 NOTICE system: link up on port 1
Jul 8 17:25:41 NOTICE system: link up on port 3
Jul 8 17:25:42 NOTICE system: link up on port 5
Jul 8 17:25:42 NOTICE system: link up on port 1
Jul 8 17:25:42 NOTICE system: link up on port 6
```

Each syslog message has a criticality level associated with it, included in text form as a prefix to the log message. One of eight different prefixes is used, depending on the condition that the administrator is being notified of:

- Level 0 - EMERG: Indicates that the system is unusable.
- Level 1 - ALERT: Indicates that action should be taken immediately.
- Level 2 - CRIT: Indicates critical conditions.
- Level 3 - ERR: Indicates error conditions or operations in error.
- Level 4 - WARNING: Indicates warning conditions.
- Level 5 - NOTICE: Indicates a normal but significant condition.
- Level 6 - INFO: Indicates an information message.
- Level 7 - DEBUG: Indicates a debug-level message.

Information logged

You can selectively choose what information should be logged by Syslog. You have number of options:

`all`	All
`bgp`	BGP
`cfg`	Configuration
`cli`	Command-line interface
`console`	Console
`dcbx`	DCB Capability Exchange
`difftrak`	Configuration difference tracking
`failover`	Failover
`fcoe`	Fibre Channel over Ethernet
`hotlinks`	Hot Links
`ip`	Internet protocol
`ipv6`	IPv6
`lacp`	Link Aggregation Control Protocol
`link`	System port link
`lldp`	LLDP
`management`	Management
`mld`	MLD
`netconf`	NETCONF Configuration Protocol
`ntp`	Network time protocol
`ospf`	OSPF
`ospfv3`	OSPFv3
`rmon`	Remote monitoring
`server`	Syslog server
`spanning-tree-group`	Spanning Tree Group
`ssh`	Secure Shell
`system`	System
`vlag`	Virtual Link Aggregation
`vlan`	VLAN
`vm`	Virtual machine
`vnic`	VNIC
`vrrp`	Virtual Router Redundancy Protocol
`web`	Web

Logging destinations

You can set up to two destinations for reporting. A destination of 0.0.0.0 means logs are stored locally on the switch. Another instance of a log destination host can be a remote logging server. In this case, the logs are sent to the server through Syslog. For each of the two destinations, you can define many parameters, including the severity of logs to be sent to that particular destination.

In Example 7-7, we set a configuration to log locally the messages with ALERT (Level 1) severity and to send all critical (severity CRIT, Level 2) events to 172.25.101.200.

Example 7-7 Example of Syslog configuration

```
RS8264(config)#logging host 1 address 0.0.0.0
RS8264(config)#logging host 1 severity 1

RS8264(config)#logging host 2 address 172.25.101.200
Jan 27 10:48:18 ACC-2 NOTICE  mgmt: second syslog host changed to 172.25.101.200
via MGTA port
RS8264(config)#logging host 2 severity 2
```

By default, logs to remote destinations are sent through the MGTA interface. You can change it to either MGTB or a data port. For example, to send the logs to a second destination from a data port, run the command shown in Example 7-7.

Example 7-8 Changing the logging interface

```
RS8264(config)#logging host 2 data-port
```

Logging console

To make logging output visible on the console, run **logging console**.

7.3.2 SNMP

IBM Networking OS provides Simple Network Management Protocol (SNMP) version 1, version 2, and version 3 support for access through any network management software, such as IBM Systems Director.

Important: SNMP read and write functions are enabled by default. If SNMP is not needed for your network, it is a preferred practice that you disable these functions before connecting the switch to the network.

SNMP versions 1 and 2

To access the SNMP agent on the RackSwitch G8264, the read and write community strings on the SNMP manager should be configured to match the community strings on the switch. The default read community string on the switch is public and the default write community string is private.

The read and write community strings on the switch can be changed by running the following commands:

```
RS8264(config)# snmp-server read-community <1-32 characters>
RS8264(config)# snmp-server write-community <1-32 characters>
```

The SNMP manager should be able to reach the management interface or any of the IP interfaces on the switch.

For the SNMP manager to receive the SNMPv1 traps sent out by the SNMP agent on the switch, configure the trap host on the switch by running the following command:

```
RS8264(config)# snmp-server trap-src-if <trap source IP interface>
RS8264(config)# snmp-server host <IPv4 address> <trap host community string>
```

SNMP version 3

SNMP version 3 (SNMPv3) is an enhanced version of the Simple Network Management Protocol, approved by the Internet Engineering Steering Group in March 2002. SNMPv3 contains additional security and authentication features that provide data origin authentication, data integrity checks, timeliness indicators, and encryption to protect against threats, such as masquerade, modification of information, message stream modification, and disclosure.

Using SNMPv3, your clients can query the MIBs securely.

Default configuration

IBM Networking OS has two SNMPv3 users by default. Both of the following users have access to all the MIBs supported by the switch:

- User 1 name is adminmd5 (password adminmd5). The authentication used is MD5.
- User 2 name is adminsha (password adminsha). The authentication used is SHA.

Up to 16 SNMP users can be configured on the switch. To modify an SNMP user, run the following command:

```
RS8264(config)# snmp-server user <1-16> name <1-32 characters>
```

Users can be configured to use the authentication and privacy options. The G8264 switch supports two authentication algorithms, MD5 and SHA, as specified in the following command:

```
RS8264(config)# snmp-server user <1-16> authentication-protocol
{md5|sha} authentication-password
```

User configuration example

To configure a user, complete the following steps:

1. To configure a user with the name `admin`, the authentication type MD5, the authentication password of `admin`, and the privacy option DES with a privacy password of `admin`, run the commands shown in Example 7-9.

Example 7-9 SNMP v3 user configuration example

```
RS8264(config)# snmp-server user 5 name admin
RS8264(config)# snmp-server user 5 authentication-protocol md5
authentication-password
Changing authentication password; validation required:
Enter current admin password: <admin. password>
Enter new authentication password: <auth. password>
Re-enter new authentication password: <auth. password>
New authentication password accepted.
RS8264(config)# snmp-server user 5 privacy-protocol des
privacy-password
Changing privacy password; validation required:
```

```
Enter current admin password: <admin. password>
Enter new privacy password: <privacy password>
Re-enter new privacy password: <privacy password>
New privacy password accepted.
```

2. Configure a user access group, along with the views the group may access, by running the commands shown in Example 7-10. Use the access table to configure the group's access level.

Example 7-10 SNMPv3 group and view configuration example

```
RS8264(config)# snmp-server access 5 name admingrp
RS8264(config)# snmp-server access 5 level authpriv
RS8264(config)# snmp-server access 5 read-view iso
RS8264(config)# snmp-server access 5 write-view iso
RS8264(config)# snmp-server access 5 notify-view iso
```

Because the read view, write view, and notify view are all set to `iso`, the user type has access to all private and public MIBs.

3. Assign the user to the user group by running the commands shown in Example 7-11. Use the group table to link the user to a particular access group.

Example 7-11 SNMPv3 user assignment configuration

```
RS8264(config)# snmp-server group 5 user-name admin
RS8264(config)# snmp-server group 5 group-name admingrp
```

Configuring SNMP traps

Here we describe the steps for configuring the SNMP traps.

SNMPv2 trap configuration

To configure the SNMPv2 trap, complete the following steps:

1. Configure a user with no authentication and password, as shown in Example 7-12.

Example 7-12 SNMP user configuration example

```
RS8264(config)#snmp-server user 10 name v2trap
```

2. Configure an access group and group table entries for the user. Use the menu shown in Example 7-13 to specify which traps can be received by the user:

Example 7-13 SNMP group configuration

```
RS8264(config)#snmp-server group 10 security snmpv2
RS8264(config)#snmp-server group 10 user-name v2trap
RS8264(config)#snmp-server group 10 group-name v2trap
RS8264(config)#snmp-server access 10 name v2trap
RS8264(config)#snmp-server access 10 security snmpv2
RS8264(config)#snmp-server access 10 notify-view iso
```

3. Configure an entry in the notify table, as shown in Example 7-14.

Example 7-14 SNMP notify entry configuration

```
RS8264(config)#snmp-server notify 10 name v2trap
RS8264(config)#snmp-server notify 10 tag v2trap
```

4. Specify the IPv4 address and other trap parameters in the targetAddr and targetParam tables. Use the commands shown in Example 7-15 to specify the user name associated with the targetParam table.

Example 7-15 SNMP trap destination and trap parameters configuration

```
RS8264(config)#snmp-server target-address 10 name v2trap address 100.10.2.1
RS8264(config)#snmp-server target-address 10 taglist v2trap
RS8264(config)#snmp-server target-address 10 parameters-name v2param
RS8264(config)#snmp-server target-parameters 10 name v2param
RS8264(config)#snmp-server target-parameters 10 message snmpv2c
RS8264(config)#snmp-server target-parameters 10 user-name v2trap
RS8264(config)#snmp-server target-parameters 10 security snmpv2
```

5. Use the community table to specify which community string is used in the trap, as shown in Example 7-16.

Example 7-16 SNMP community configuration

```
RS8264(config)#snmp-server community 10 index v2trap
RS8264(config)#snmp-server community 10 user-name v2trap
```

SNMPv3 trap configuration

To configure a user for SNMPv3 traps, you can choose to send the traps with both privacy and authentication, with authentication only, or without privacy or authentication.

You can configure these settings in the access table by running the following commands:

- **RS8264(config)#snmp-server access <1-32> level**
- **RS8264(config)#snmp-server target-parameters <1-16>**

Configure the user in the user table.

It is not necessary to configure the community table for SNMPv3 traps, because the community string is not used by SNMPv3.

Example 7-17 shows how to configure a SNMPv3 user v3trap with authentication only:

Example 7-17 SNMPv3 trap configuration

```
RS8264(config)#snmp-server user 11 name v3trap
RS8264(config)#snmp-server user 11 authentication-protocol md5
authentication-password
Changing authentication password; validation required:
Enter current admin password: <admin. password>
Enter new authentication password: <auth. password>
Re-enter new authentication password: <auth. password>
New authentication password accepted.
RS8264(config)#snmp-server access 11 notify-view iso
RS8264(config)#snmp-server access 11 level authnopriv
RS8264(config)#snmp-server group 11 user-name v3trap
RS8264(config)#snmp-server group 11 tag v3trap
RS8264(config)#snmp-server notify 11 name v3trap
RS8264(config)#snmp-server notify 11 tag v3trap
RS8264(config)#snmp-server target-address 11 name v3trap address
172.25.101.200
RS8264(config)#snmp-server target-address 11 taglist v3trap
RS8264(config)#snmp-server target-address 11 parameters-name v3param
```

```
RS8264(config)#snmp-server target-parameters 11 name v3param
RS8264(config)#snmp-server target-parameters 11 user-name v3trap
RS8264(config)#snmp-server target-parameters 11 level authNoPriv
```

7.3.3 Remote Monitoring (RMON)

The IBM switches provide a Remote Monitoring (RMON) interface that allows network devices to exchange network monitoring data. RMON allows the switch to perform the following functions:

- Track events and trigger alarms when a threshold is reached.
- Notify administrators by issuing a syslog message or SNMP trap.

The RMON MIB provides an interface between the RMON agent on the switch and an RMON management application. The RMON MIB is described in RFC 1757 (http://www.ietf.org/rfc/rfc1757.txt). The RMON standard defines objects that are suitable for the management of Ethernet networks. The RMON agent continuously collects statistics and proactively monitors switch performance. You can use RMON to monitor traffic that flows through the switch.

The switch supports the following RMON Groups, as described in RFC 1757:

- Group 1: Statistics
- Group 2: History
- Group 3: Alarms
- Group 9: Events

RMON Group 1: Statistics

The switch supports collection of Ethernet statistics as outlined in the RMON statistics MIB, referring to etherStatsTable. You can configure RMON statistics on a per-port basis. RMON statistics are sampled every second, and new data overwrites any old data on a port.

Important: RMON port statistics must be enabled for the port before you can view RMON statistics.

Example configuration

Here is an example configuration:

1. Enable RMON on a port. To enable RMON on a port, run **interface** and **rmon**:
   - **RS8264(config)# interface port 1**
   - **RS8264(config-if)# rmon**
2. To view the RMON statistics, run **interface**, run **rmon**, and run **show** to show the inteface, as shown in Example 7-18.

Example 7-18 View of the RMON statistics

```
RS8264(config)# interface port 1
RS8264(config-if)# rmon
RS8264(config-if)# show interface port 1 rmon-counters
----------------------------------------------------------------
RMON statistics for port 3:
etherStatsDropEvents: NA
etherStatsOctets: 7305626
etherStatsPkts: 48686
```

```
etherStatsBroadcastPkts: 4380
etherStatsMulticastPkts: 6612
etherStatsCRCAlignErrors: 22
etherStatsUndersizePkts: 0
etherStatsOversizePkts: 0
etherStatsFragments: 2
etherStatsJabbers: 0
etherStatsCollisions: 0
etherStatsPkts64Octets: 27445
etherStatsPkts65to127Octets: 12253
etherStatsPkts128to255Octets: 1046
etherStatsPkts256to511Octets: 619
etherStatsPkts512to1023Octets: 7283
etherStatsPkts1024to1518Octets: 38
```

RMON Group 2: History

You can use the RMON History Group to sample and archive Ethernet statistics for a specific interface during a specific time interval. History sampling is done per port.

Important: RMON port statistics must be enabled for the port before an RMON History Group can monitor the port.

Data is stored in buckets, which store data gathered during discreet sampling intervals. At each configured interval, the History index takes a sample of the current Ethernet statistics, and places them into a bucket. History data buckets are in dynamic memory. When the switch is rebooted, the buckets are emptied.

Requested buckets are the number of buckets, or data slots, requested by the user for each History Group. Granted buckets are the number of buckets granted by the system, based on the amount of system memory available. The system grants a maximum of 50 buckets.

You can use an SNMP browser to view History samples.

History MIB Object ID

The type of data that can be sampled must be of an ifIndex object type, as described in RFC 1213 (http://www.ietf.org/rfc/rfc1213.txt) and RFC 1573 (http://www.ietf.org/rfc/rfc1573.txt). The most common data type for the History sample is as follows:

`1.3.6.1.2.1.2.2.1.1.<x>`

The last digit (x) represents the number of the port to monitor.

7.3.4 Management applications

After you set up the logs, either with Syslog or SNMP, you can now receive and browse them in Network Management System or any monitoring application. Because IBM switches use standard protocols to format and send the logs, you can access them with any management software. In this section, we how to access logs from your devices from within IBM System Networking Element Manager and IBM Systems Director software.

IBM System Networking Element Manager

IBM System Networking Element Manager (SNEM) is an application for remote monitoring and management of Ethernet switches from IBM. It is designed to simplify and centralize the management of your BladeCenter or blade server and Top-of-Rack Ethernet switches. IBM System Networking Element Manager is a simple yet powerful tool that you can use to easily set up a logging and reporting environment to monitor the devices from a central point.

The SNEM home page gives a quick summary of the discovered devices. It provides a graphical representation of Health Status, Panic Dump, Events, Save Pending, Running Software Version, and Device Discovery Timestamp, grouped into separate panes along with the device counts.

A sample System Networking Element Manager home page is shown in Figure 7-1.

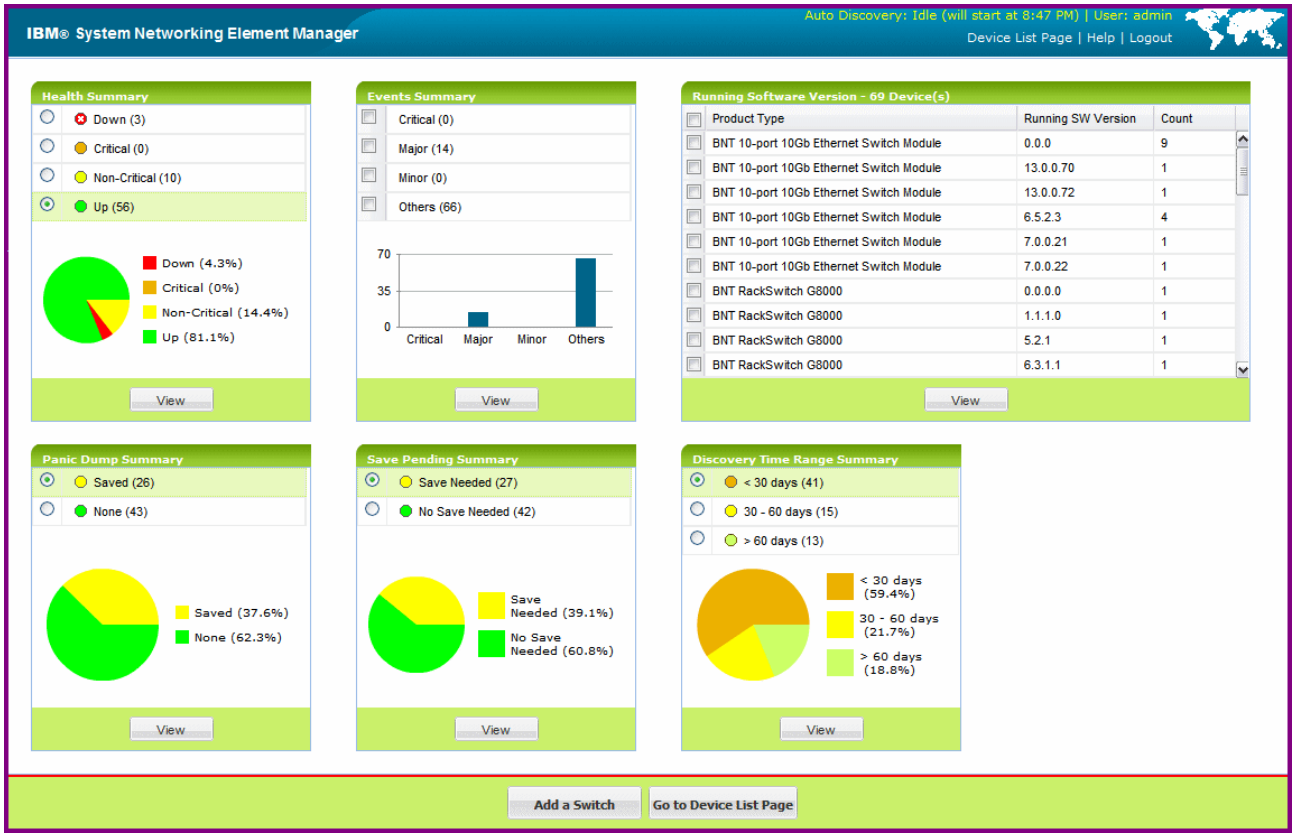

Figure 7-1 System Networking Element Manager home page example

The information is updated periodically to provide the actual counts and status of managed devices. It provides an option for the user to filter the devices available on the device list page based on the selection made here.

Health Status Summary pane

The Health Status Summary pane shows the individual count of devices discovered that are Down (red), Critical (orange), Non-Critical (yellow), and Up (green). It also provides a pie chart that indicates the percentages of Down/Critical/Non-Critical/Up devices. A sample Health Status Summary pane is shown in Figure 7-2.

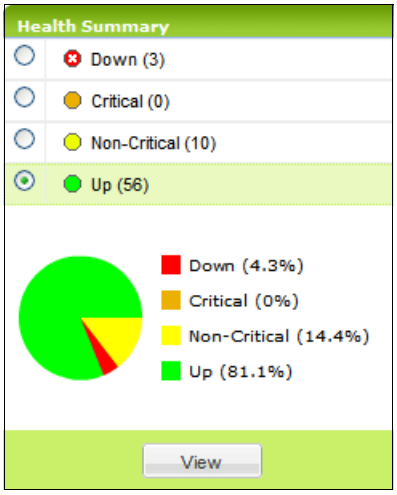

Figure 7-2 Health Summary pane

You can filter out the devices of the Health Status by selecting the appropriate choice and clicking **View**, which takes you to the Device list page.

Viewing Health Status

The Health Status page shows processor and Memory Utilization, ARP and Routing Table Utilization, Power Supply Status, Panic Dump Status, Temperature Sensors reading, and Fan Speed. A sample Health Status window is shown in Figure 7-3.

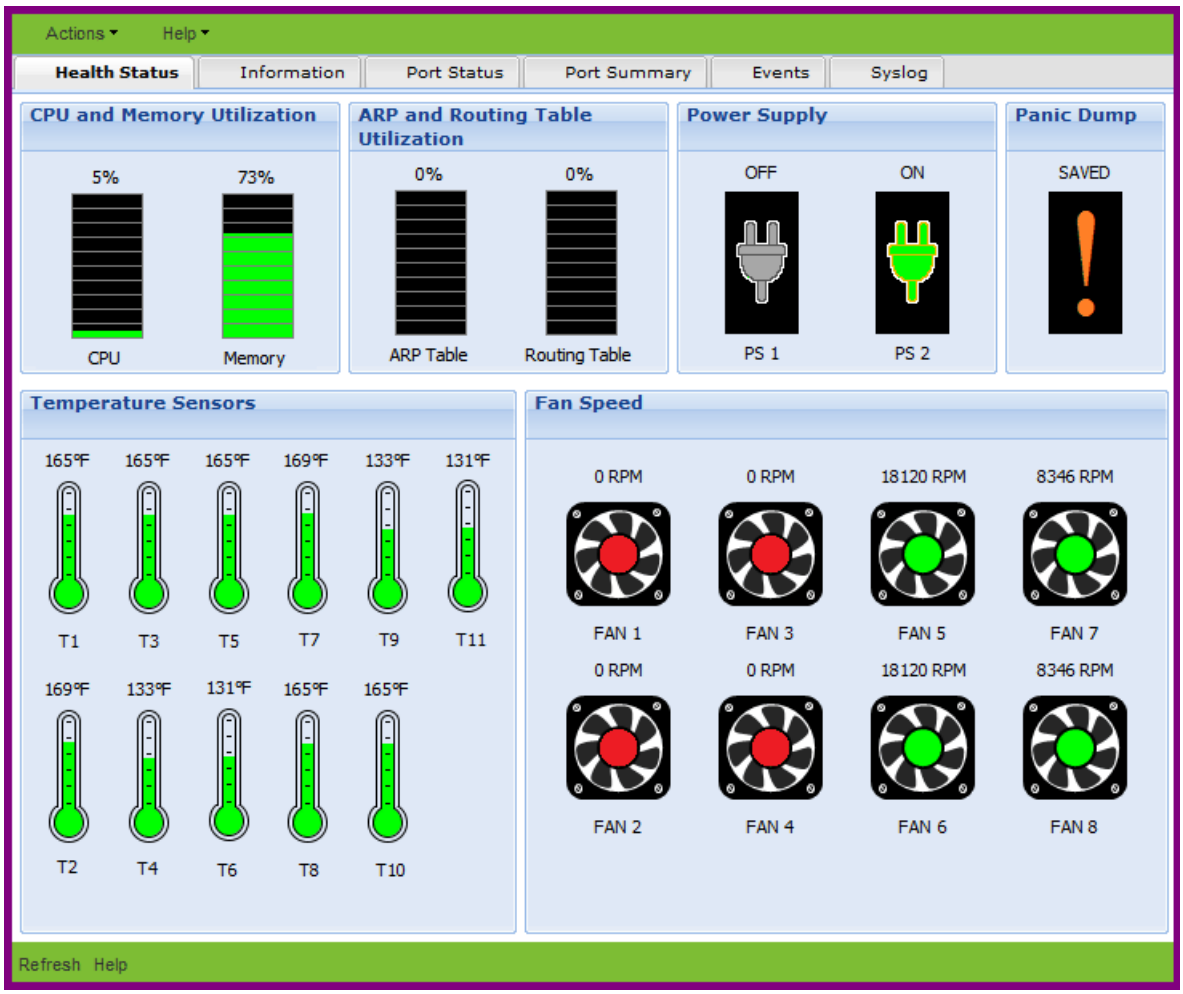

Figure 7-3 Health Status window

Viewing reports

You can view various reports associated with all the discovered switches by choosing the items under the Reports menu in SNEM.

The list of reports include:

- Event List Report
- Syslog List Report
- SNEM Alerts Report
- Switch Version Report
- Transceiver Information Report
- VM Data Center Report
- VMready VM Report
- VMready VM Report – Port Groups Report

For more information about SNEM, see the following publications:

- *IBM SNEM 6.1 Solution Getting Started Guide*:

  http://www-01.ibm.com/support/docview.wss?uid=isg3T7000471&aid=1
- *IBM SNEM 6.1 User Guide*:

  http://www-01.ibm.com/support/docview.wss?uid=isg3T7000473&aid=1
- *IBM SNEM 6.1 Release Notes Changes*:

  http://www-01.ibm.com/support/docview.wss?uid=isg3T7000474&aid=1
- *IBM System Networking Element Manager Solution Device Support List (6.1)*:

  http://www-01.ibm.com/support/docview.wss?uid=isg3T7000474
- *Quick Start Guide for installing and running KVM*:

  http://publib.boulder.ibm.com/infocenter/lnxinfo/v3r0m0/topic/liaai/kvminstall/kvminstall_pdf.pdf

IBM Systems Director

IBM Systems Director is another application from IBM that can be used by IBM System Networking switches as a logging and reporting destination. For more information about IBM System Director, see *Implementing IBM Systems Director 6.1*, SG24-7694.

7.4 Troubleshooting

In this section, we show some basic troubleshooting tools and techniques.

7.4.1 LEDs

You can verify the basic status of the switch by visually inspecting the switch's LEDs.

For information about the LED statuses of different switch models, see the LED Status sections of Chapter 1, "Introduction to IBM System Networking 10Gb Ethernet products" on page 1.

7.4.2 Basic troubleshooting procedure

This section contains basic troubleshooting information to help resolve problems that might occur during the installation and operation of your switch. If you have problems accessing the switch or working with the software, see the Command Reference guide for your switch model, as listed in "Related publications" on page 333.

System LEDs do not light

Symptom: The Power Supply LED does not light.

Solution: Check the power supply to make sure that there is a connection to the power source. Verify that power is available from the power source.

Port link LED does not light

Symptom: The port link LED does not light.

Solution 1: Check the port configuration in the software (see the Command Reference for your switch). If the port is configured with a specific speed or duplex mode, check the other device to verify that it is set to the same configuration. If the switch port is set to autonegotiate, verify that the other device is set to autonegotiate.

Solution 2: Check the cables that connect the port to the other device. Make sure that they are connected. Verify that you are using the correct cable type.

Temperature sensor warning

Symptom: A temperature warning shows on the management console.

Solution: Make sure that the air circulation vents on the front, back, and sides of the switch are free from obstruction by cables, panels, rack frames, or other materials.

Make sure that all cooling fans inside the switch are running. A Fan Module's LED (rear panel) flashes if there is a failure of the fan. Run the following command to show the fan status:

```
show sys-info
```

If any fan stops during switch operation, contact IBM Support.

Switch does not initialize (boot)

Symptom: All the switch LEDs stay on, and the command prompt does not appear on the console.

Solution: The operating system might be damaged. Use the console port to perform a serial upgrade of the switch software. See the Command Reference guide for you switch, as listed in "Related publications" on page 333.

7.4.3 Connectivity troubleshooting

In this section, you find basic information about how to troubleshoot the IP connectivity in a network built on IBM System Networking switches. IBM switches come with a set of simple tools that can be helpful for troubleshooting IP connectivity issues.

Ping

The **`ping`** command is a simple tool, based on a request-response mechanism, to verify connectivity to a remote network node. The **`ping`** command is based on ICMP. The request is an ICMP Echo packet and the reply is an ICMP Echo Reply. Like a regular IP packet, an ICMP packet is forwarded based on the intermediate routers' routing table until it reaches the destination. After it reaches the destination, the ICMP Echo Reply packet is generated and forwarded back to the originating node.

Important: In IBM switches, **`ping`** sends an ICMP Echo packet out of the management interface first. If you want to change that option, you need to add the `data-port` keyword to a command as a parameter.

For example, to verify the connectivity from the ACC-2 switch used in Chapter 3, "Reference architectures" on page 107 to the AGG-1 switch's VLAN100 IP address (10.0.100.1), run the command shown in Example 7-19.

Example 7-19 Ping command example

```
ACC-2#ping 10.0.100.1 data-port
```

```
Connecting via DATA port.
[host 10.0.100.1, max tries 5, delay 1000 msec , length 0]
10.0.100.1: #1 ok, RTT 0 msec.
10.0.100.1: #2 ok, RTT 0 msec.
10.0.100.1: #3 ok, RTT 1 msec.
10.0.100.1: #4 ok, RTT 0 msec.
10.0.100.1: #5 ok, RTT 0 msec.
```

You can see in the output that all five ICMP Echo requests received the replies. There is also additional information about the Round Trip Time (RTT), that is, the time it took for the ACC-2 to receive the response from AGG-1. `0 msec` means that the time was less than 1 ms.

Traceroute

You can use the **`traceroute`** command to not only verify connectivity to a remote network node, but to track the responses from intermediate nodes as well. This action is done by using the Time-To-Live (TTL) field in IP packets. The **`traceroute`** command sends a UDP packet to a port that is likely to not be used on a remote node with a TTL of 1. After the packet reaches the intermediate router, the TTL is decremented, and the ICMP time-exceeded message is sent back to the originating node, which increments the TTL to 2, and the process repeats. After the UDP packet reaches a destination host, an ICMP port-unreachable message is sent back to the sender. This action provides the sender with information about all intermediate routers on the way to the destination.

The command shown in Example 7-20 verifies which hops are on the way from the ACC-2 switch to AGG-1.

Example 7-20 Traceroute command example

```
ACC-2#traceroute 10.0.100.1 data-port
Connecting via DATA port.
[host 10.0.100.1, max-hops 32, delay 2048 msec]
 1  10.0.100.1       0 ms
Trace host responded.
```

From the output, you see that there is only one hop on the way from ACC-2 to AGG-1, and it is AGG-1 itself. We use OSPF in our network, which selects this path as the shortest one.

For test purposes, we shut down the direct links between ACC-2 and AGG-1 (ports 3 and 4) and run **`traceroute`** again. The output is shown in Example 7-21.

Example 7-21 Traceroute command example

```
ACC-2#traceroute 10.0.100.1 data-port
Connecting via DATA port.
[host 10.0.100.1, max-hops 32, delay 2048 msec]
 1  10.0.104.1       0 ms
 2  10.0.100.1       1 ms
Trace host responded.
```

Now we can see that in order to get to AGG-1, ACC-2 uses the AGG-2 (`10.0.104.1`) switch as the intermediate router.

A

Configuration files

This appendix provides the final working configuration of the equipment used for the reference architecture. The configuration files are from the following equipment:

- AGG-1: Aggregation switch (RackSwitch G8264)
- AGG-2: Aggregation switch (RackSwitch G8264)
- ACC-1: Access switch (RackSwitch G8124)
- ACC-2: Access switch (RackSwitch G8124)
- ACC-3: Access switch (Virtual Fabric 10Gb Switch Module stack)

© Copyright IBM Corp. 2012. All rights reserved.

AGG-1: Aggregation switch (RackSwitch G8264)

Example A-1 shows the final configuration of the AGG-1 aggregation switch.

Example A-1 Final configuration of the AGG-1 aggregation switch

```
AGG-1#
!
version "6.8.0.3"
switch-type "Blade Network Technologies RackSwitch G8264"
!
!
ssh scp-enable
ssh enable
!
!
!
no system bootp
no system dhcp
hostname "AGG-1"
system idle 60
!
!
interface port 1
        name "AGG1-AGG2"
        tagging
        pvid 100
        exit
!
interface port 5
        name "AGG1-AGG2"
        tagging
        pvid 100
        exit
!
interface port 17
        name "AGG1-ACC1"
        pvid 101
        exit
!
interface port 18
        name "AGG1-ACC1"
        pvid 101
        exit
!
interface port 19
        name "AGG1-ACC2"
        pvid 103
        exit
!
interface port 20
        name "AGG1-ACC2"
        pvid 103
        exit
!
```

```
interface port 21
        name "AGG1-ACC3"
        pvid 30
        exit
!
interface port 22
        name "AGG1-AGG4"
        pvid 30
        exit
!
vlan 1
        member 2-4,6-16,23-64
        no member 1,5,17-22
!
vlan 30
        enable
        name "SRV3"
        member 1,5,21-22
!
vlan 100
        enable
        name "IPv4_AGG1-AGG2"
        member 1,5
!
vlan 101
        enable
        name "IPv4_AGG1-ACC1"
        member 17-18
!
vlan 103
        enable
        name "IPv4_AGG1-ACC2"
        member 19-20
!
portchannel 1 port 17
portchannel 1 port 18
portchannel 1 enable
!
portchannel 2 port 19
portchannel 2 port 20
portchannel 2 enable
!
portchannel 3 port 1
portchannel 3 port 5
portchannel 3 enable
!
portchannel 4 port 21
portchannel 4 port 22
portchannel 4 enable
!
portchannel thash l2thash l2-source-mac-address
!
portchannel thash ingress
!
!
```

```
spanning-tree stp 30 bridge priority 0
spanning-tree stp 30 vlan 30

spanning-tree stp 100 bridge priority 0
spanning-tree stp 100 vlan 100

spanning-tree stp 101 bridge priority 0
spanning-tree stp 101 vlan 101

spanning-tree stp 103 bridge priority 0
spanning-tree stp 103 vlan 103

!
logging host 2 address 10.10.53.219 MGT
!
lldp enable
!
ip router-id 1.1.1.1
!
interface ip 30
        ip address 10.0.30.2 255.255.255.0
        vlan 30
        enable
        exit
!
interface ip 36
        ipv6 address fc30:0:0:0:0:0:0:2 64
        ipv6 secaddr6 address fc30:0:0:0:0:0:0:1 64 anycast
        vlan 30
        enable
        no ipv6 nd suppress-ra
        ipv6 nd prefix fc30:0:0:0:0:0:0:0 64
        exit
!
interface ip 100
        ip address 10.0.100.1 255.255.255.252
        vlan 100
        enable
        exit
!
interface ip 101
        ip address 10.0.101.2 255.255.255.252
        vlan 101
        enable
        exit
!
interface ip 103
        ip address 10.0.103.1 255.255.255.252
        vlan 103
        enable
        exit
!
interface ip 110
        ipv6 address fc00:0:0:0:0:0:0:1 64
        vlan 100
```

```
        enable
        ip6host
        exit
!
interface ip 111
        ipv6 address fc11:0:0:0:0:0:0:2 64
        vlan 101
        enable
        ip6host
        exit
!
interface ip 113
        ipv6 address fc13:0:0:0:0:0:0:1 64
        vlan 103
        enable
        ip6host
        exit
!
interface ip 128
        ip address 172.25.101.120
        enable
        exit
!
interface loopback 1
        ip address 1.1.1.1  255.255.255.255
        enable
        exit
!
ip gateway 4 address 172.25.1.1
ip gateway 4 enable
!
!
!
!
!
!
router vrrp
        enable
!
        tracking-priority-increment ports 50
!
        virtual-router 3 virtual-router-id 30
        virtual-router 3 interface 30
        virtual-router 3 priority 105
        virtual-router 3 address 10.0.30.1
        virtual-router 3 enable
        virtual-router 3 timers advertise 2
        virtual-router 3 timers preempt-delay-time 5
        virtual-router 3 track ports
!
router ospf
        enable
!
        area 0 authentication-type md5
        area 0 enable
```

```
!
        redistribute fixed export 2 1
!
        message-digest-key 1 md5-ekey
ea28046b4028002ab376e7a28398a3d86e6f385d01eb3bc9926a86856348faa4bcfb83951d8c8d2fee
026eccb994eb6189d271a8be987bb684edb91152d0b937
!
interface ip 100
        ip ospf enable
        ip ospf message-digest-key 1
!
interface ip 101
        ip ospf enable
        ip ospf message-digest-key 1
!
interface ip 103
        ip ospf enable
        ip ospf message-digest-key 1
!
ipv6 router ospf
        router-id 1.1.1.1
        enable
!
        area 0 area-id 0.0.0.0
        area 0 stability-interval 40
        area 0 default-metric 1
        area 0 default-metric type 1
        area 0 translation-role candidate
        area 0 type transit
        area 0 enable
!
        redistribute connected export 2 1
!
interface ip 110
        ipv6 ospf area 0
        ipv6 ospf retransmit-interval 5
        ipv6 ospf transmit-delay 1
        ipv6 ospf priority 1
        ipv6 ospf hello-interval 10
        ipv6 ospf dead-interval 40
        ipv6 ospf cost 1
        ipv6 ospf enable
!
interface ip 111
        ipv6 ospf area 0
        ipv6 ospf retransmit-interval 5
        ipv6 ospf transmit-delay 1
        ipv6 ospf priority 1
        ipv6 ospf hello-interval 10
        ipv6 ospf dead-interval 40
        ipv6 ospf cost 1
        ipv6 ospf enable
!
interface ip 113
        ipv6 ospf area 0
```

```
        ipv6 ospf retransmit-interval 5
        ipv6 ospf transmit-delay 1
        ipv6 ospf priority 1
        ipv6 ospf hello-interval 10
        ipv6 ospf dead-interval 40
        ipv6 ospf cost 1
        ipv6 ospf enable
!
end
```

AGG-2: Aggregation switch (RackSwitch G8264)

Example A-2 shows the final configuration of the AGG-2 aggregation switch.

Example A-2 Final configuration of the AGG-2 aggregation switch

```
version "6.8.0.3"
switch-type "Blade Network Technologies RackSwitch G8264"
!
!
ssh scp-enable
ssh enable
!
!
!
no system bootp
no system dhcp
hostname "AGG-2"
system idle 60
!
!
interface port 1
        name "AGG1-AGG2"
        tagging
        pvid 100
        exit
!
interface port 5
        name "AGG1-AGG2"
        tagging
        pvid 100
        exit
!
interface port 17
        name "AGG2-ACC2"
        pvid 104
        exit
!
interface port 18
        name "AGG2-ACC2"
        pvid 104
        exit
!
interface port 19
```

```
        name "AGG2-ACC1"
        pvid 102
        exit
!
interface port 20
        name "AGG2-ACC1"
        pvid 102
        exit
!
interface port 21
        name "AGG2-ACC3"
        pvid 30
        exit
!
interface port 22
        name "AGG2-AGG4"
        pvid 30
        exit
!
vlan 1
        member 2-4,6-16,23-64
        no member 1,5,17-22
!
vlan 30
        enable
        name "IPv4_SRV3"
        member 1,5,21-22
!
vlan 100
        enable
        name "AGG1-AGG2"
        member 1,5
!
vlan 102
        enable
        name "AGG2-ACC1"
        member 19-20
!
vlan 104
        enable
        name "AGG2-ACC2"
        member 17-18
!
portchannel 1 port 17
portchannel 1 port 18
portchannel 1 enable
!
portchannel 2 port 19
portchannel 2 port 20
portchannel 2 enable
!
portchannel 3 port 1
portchannel 3 port 5
portchannel 3 enable
!
```

```
portchannel 4 port 21
portchannel 4 port 22
portchannel 4 enable
!
portchannel thash l2thash l2-source-mac-address
!
portchannel thash ingress
!
!
spanning-tree stp 30 bridge priority 4096
spanning-tree stp 30 vlan 30

spanning-tree stp 100 vlan 100

spanning-tree stp 102 vlan 102

spanning-tree stp 104 vlan 104

!
!
lldp enable
!
ip router-id 1.1.1.2
!
interface ip 30
        ip address 10.0.30.3 255.255.255.0
        vlan 30
        enable
        exit
!
interface ip 36
        ipv6 address fc30:0:0:0:0:0:0:3 64
        ipv6 secaddr6 address fc30:0:0:0:0:0:0:1 64 anycast
        vlan 30
        enable
        exit
!
interface ip 100
        ip address 10.0.100.2 255.255.255.252
        vlan 100
        enable
        exit
!
interface ip 102
        ip address 10.0.102.2 255.255.255.252
        vlan 102
        enable
        exit
!
interface ip 104
        ip address 10.0.104.1 255.255.255.252
        vlan 104
        enable
        exit
!
```

```
interface ip 110
        ipv6 address fc00:0:0:0:0:0:0:2 64
        vlan 100
        enable
        ip6host
        exit
!
interface ip 112
        ipv6 address fc12:0:0:0:0:0:0:2 64
        vlan 102
        enable
        ip6host
        exit
!
interface ip 114
        ipv6 address fc14:0:0:0:0:0:0:1 64
        vlan 104
        enable
        ip6host
        exit
!
interface ip 128
        ip address 172.25.101.121 255.255.255.0
        enable
        exit
!
interface loopback 1
        ip address 1.1.1.2  255.255.255.255
        enable
        exit
!
ip gateway 4 address 172.25.101.1
ip gateway 4 enable
!
!
!
!
!
!
router vrrp
        enable
!
        tracking-priority-increment ports 50
!
        virtual-router 3 virtual-router-id 30
        virtual-router 3 interface 30
        virtual-router 3 address 10.0.30.1
        virtual-router 3 enable
        virtual-router 3 timers advertise 2
        virtual-router 3 timers preempt-delay-time 5
        virtual-router 3 track ports
!
router ospf
        enable
!
```

```
        area 0 authentication-type md5
        area 0 enable
!
        redistribute fixed export 5 1
!
        message-digest-key 1 md5-ekey
2d6914f4002000a024a0f7b7c390a3528f55837641b160226c2c717671830ebc1d2e3de38fc850c1e0
280022ea6d7f87350fe239c7568803d9090116b2482cc0
!
interface ip 100
        ip ospf enable
        ip ospf message-digest-key 1
!
interface ip 102
        ip ospf enable
        ip ospf message-digest-key 1
!
interface ip 104
        ip ospf enable
        ip ospf message-digest-key 1
!
ipv6 router ospf
        router-id 1.1.1.2
        enable
!
        area 0 area-id 0.0.0.0
        area 0 stability-interval 40
        area 0 default-metric 1
        area 0 default-metric type 1
        area 0 translation-role candidate
        area 0 type transit
        area 0 enable
!
        redistribute connected export 5 1
!
interface ip 110
        ipv6 ospf area 0
        ipv6 ospf retransmit-interval 5
        ipv6 ospf transmit-delay 1
        ipv6 ospf priority 1
        ipv6 ospf hello-interval 10
        ipv6 ospf dead-interval 40
        ipv6 ospf cost 1
        ipv6 ospf enable
!
interface ip 112
        ipv6 ospf area 0
        ipv6 ospf retransmit-interval 5
        ipv6 ospf transmit-delay 1
        ipv6 ospf priority 1
        ipv6 ospf hello-interval 10
        ipv6 ospf dead-interval 40
        ipv6 ospf cost 1
        ipv6 ospf enable
!
```

```
interface ip 114
        ipv6 ospf area 0
        ipv6 ospf retransmit-interval 5
        ipv6 ospf transmit-delay 1
        ipv6 ospf priority 1
        ipv6 ospf hello-interval 10
        ipv6 ospf dead-interval 40
        ipv6 ospf cost 1
        ipv6 ospf enable
!
end
```

ACC-1: Access switch (RackSwitch G8124)

Example A-3 shows the final configuration of the ACC-1 access switch.

Example A-3 Final configuration of the ACC-1 access switch

```
version "6.8.0.3"
switch-type "Blade Network Technologies RackSwitch G8124-E"
!
!
ssh scp-enable
ssh enable
!
!
!
no system bootp
no system dhcp mgta
hostname "ACC-1"
system idle 60
!
!
interface port 1
        name "AGG1-ACC1"
        pvid 101
        exit
!
interface port 2
        name "AGG1-ACC1"
        pvid 101
        exit
!
interface port 3
        name "AGG2-ACC1"
        pvid 102
        exit
!
interface port 4
        name "AGG2-ACC1"
        pvid 102
        exit
!
interface port 5
```

```
        name "ACC1-ACC2"
        tagging
        pvid 10
        exit
!
interface port 6
        name "ACC1-ACC2"
        tagging
        pvid 10
        exit
!
interface port 7
        name "SRV1"
        pvid 10
        exit
!
vlan 1
        member 8-24
        no member 1-7
!
vlan 10
        enable
        name "IPv4_SRV1"
        member 5-7
!
vlan 101
        enable
        name "IPv4_ACC1-AGG1"
        member 1-2
!
vlan 102
        enable
        name "IPv4_ACC1-AGG2"
        member 3-4
!
portchannel 1 port 1
portchannel 1 port 2
portchannel 1 enable
!
portchannel 2 port 3
portchannel 2 port 4
portchannel 2 enable
!
portchannel hash source-mac-address
!
!
spanning-tree stp 10 bridge priority 0
spanning-tree stp 10 vlan 10

spanning-tree stp 101 vlan 101

spanning-tree stp 102 vlan 102

!
interface port 5
```

```
        lacp mode active
        lacp priority 16384
        lacp key 3
!
interface port 6
        lacp mode active
        lacp key 3
!
failover enable
failover trigger 1 limit 2
failover trigger 1 mmon monitor PortChannel 1
failover trigger 1 mmon monitor PortChannel 2
failover trigger 1 mmon control member 7
failover trigger 1 enable
!
!
!
!
!
!
!
!
!
lldp enable
!
ip router-id 2.2.2.1
!
interface ip 10
        ip address 10.0.10.2 255.255.255.0
        vlan 10
        enable
        exit
!
interface ip 101
        ip address 10.0.101.1 255.255.255.252
        vlan 101
        enable
        exit
!
interface ip 102
        ip address 10.0.102.1 255.255.255.252
        vlan 102
        enable
        exit
!
interface ip 106
        ipv6 address fc10:0:0:0:0:0:0:2 64
        ipv6 secaddr6 address fc10:0:0:0:0:0:0:1 64 anycast
        vlan 10
        enable
        ipv6 nd prefix fc10:0:0:0:0:0:0:0 64
        exit
!
interface ip 111
        ipv6 address fc11:0:0:0:0:0:0:1 64
```

```
        vlan 101
        enable
        ip6host
        exit
!
interface ip 112
        ipv6 address fc12:0:0:0:0:0:0:1 64
        vlan 102
        enable
        ip6host
        exit
!
interface ip 127
        ip address 172.25.101.122
        enable
        exit
!
interface loopback 1
        ip address 2.2.2.1  255.255.255.255
        enable
        exit
!
ip gateway 3 address 172.25.1.1
ip gateway 3 enable
!
!
!
!
!
!
ip route healthcheck
ip route ecmphash protocol
!
router vrrp
        enable
!
        tracking-priority-increment ports 50
!
        virtual-router 1 virtual-router-id 10
        virtual-router 1 interface 10
        virtual-router 1 priority 105
        virtual-router 1 address 10.0.10.1
        virtual-router 1 enable
        virtual-router 1 timers advertise 2
        virtual-router 1 timers preempt-delay-time 5
        virtual-router 1 track ports
!
router ospf
        enable
!
        area 0 authentication-type md5
        area 0 enable
!
        redistribute fixed export 2 1
!
```

```
        message-digest-key 1 md5-ekey
28e2cc3e0862882af1a3a7f7cbd22bd8f4046e5ad324465d9b1565b6fce72e2e7feafe9495d94195ae
cb3a9ef27493713d8790e864829b90bc64ce5ffaaef852
!
interface ip 101
        ip ospf enable
        ip ospf message-digest-key 1
!
interface ip 102
        ip ospf enable
        ip ospf message-digest-key 1
!
ipv6 router ospf
        router-id 2.2.2.1
        enable
!
        area 0 area-id 0.0.0.0
        area 0 stability-interval 40
        area 0 default-metric 1
        area 0 default-metric type 1
        area 0 translation-role candidate
        area 0 type transit
        area 0 enable
!
        redistribute connected export 2 1
!
interface ip 111
        ipv6 ospf area 0
        ipv6 ospf retransmit-interval 5
        ipv6 ospf transmit-delay 1
        ipv6 ospf priority 1
        ipv6 ospf hello-interval 10
        ipv6 ospf dead-interval 40
        ipv6 ospf cost 1
        ipv6 ospf enable
!
interface ip 112
        ipv6 ospf area 0
        ipv6 ospf retransmit-interval 5
        ipv6 ospf transmit-delay 1
        ipv6 ospf priority 1
        ipv6 ospf hello-interval 10
        ipv6 ospf dead-interval 40
        ipv6 ospf cost 1
        ipv6 ospf enable
!
end
```

ACC-2: Access switch (RackSwitch G8124)

Example A-4 shows the final configuration of the ACC-2 access switch.

Example A-4 Final configuration of the ACC-2 access switch

```
version "6.8.0.3"
switch-type "Blade Network Technologies RackSwitch G8124-E"
!
!
ssh scp-enable
ssh enable
!
!
!
no system bootp
no system dhcp mgta
hostname "ACC-2"
system idle 60
!
!
interface port 1
        name "AGG2-ACC2"
        pvid 104
        exit
!
interface port 2
        name "AGG2-ACC2"
        pvid 104
        exit
!
interface port 3
        name "AGG1-ACC2"
        pvid 103
        exit
!
interface port 4
        name "AGG1-ACC2"
        pvid 103
        exit
!
interface port 5
        name "ACC1-ACC2"
        tagging
        pvid 10
        exit
!
interface port 6
        name "ACC1-ACC2"
        tagging
        pvid 10
        exit
!
interface port 7
        name "SRV1"
```

```
        pvid 10
        exit
!
vlan 1
        member 8-24
        no member 1-7
!
vlan 10
        enable
        name "IPv4_SRV1"
        member 5-7
!
vlan 103
        enable
        name "IPv4_ACC2-AGG1"
        member 3-4
!
vlan 104
        enable
        name "IPv4_ACC2-AGG2"
        member 1-2
!
portchannel 1 port 1
portchannel 1 port 2
portchannel 1 enable
!
portchannel 2 port 3
portchannel 2 port 4
portchannel 2 enable
!
portchannel hash source-mac-address
!
!
spanning-tree stp 10 bridge priority 4096
spanning-tree stp 10 vlan 10

spanning-tree stp 103 vlan 103

spanning-tree stp 104 vlan 104

!
interface port 5
        lacp mode active
        lacp priority 16384
        lacp key 3
!
interface port 6
        lacp mode active
        lacp key 3
!
failover enable
failover trigger 1 limit 2
failover trigger 1 mmon monitor PortChannel 1
failover trigger 1 mmon monitor PortChannel 2
failover trigger 1 mmon control member 7
```

```
failover trigger 1 enable
!
!
!
!
!
!
!
!
!
lldp enable
!
ip router-id 2.2.2.2
!
interface ip 10
        ip address 10.0.10.3 255.255.255.0
        vlan 10
        enable
        exit
!
interface ip 103
        ip address 10.0.103.2 255.255.255.252
        vlan 103
        enable
        exit
!
interface ip 104
        ip address 10.0.104.2 255.255.255.252
        vlan 104
        enable
        exit
!
interface ip 106
        ipv6 address fc10:0:0:0:0:0:0:3 64
        ipv6 secaddr6 address fc10:0:0:0:0:0:0:1 64 anycast
        vlan 10
        enable
        exit
!
interface ip 113
        ipv6 address fc13:0:0:0:0:0:0:2 64
        vlan 103
        enable
        ip6host
        exit
!
interface ip 114
        ipv6 address fc14:0:0:0:0:0:0:2 64
        vlan 104
        enable
        ip6host
        exit
!
interface ip 127
        ip address 172.25.101.123
```

```
        enable
        exit
!
interface loopback 1
        ip address 2.2.2.2  255.255.255.255
        enable
        exit
!
ip gateway 3 address 172.25.1.1
ip gateway 3 enable
!
!
router vrrp
        enable
!
        tracking-priority-increment ports 50
!
        virtual-router 1 virtual-router-id 10
        virtual-router 1 interface 10
        virtual-router 1 address 10.0.10.1
        virtual-router 1 enable
        virtual-router 1 timers advertise 2
        virtual-router 1 timers preempt-delay-time 5
        virtual-router 1 track ports
!
router ospf
        enable
!
        area 0 authentication-type md5
        area 0 enable
!
        redistribute fixed export 2 1
!
        message-digest-key 1 md5-ekey
d3980da6519008a292b3e6e79220ab50f3d4c0a89f92b15c3f9248625f52e02b99460a33643c9330fd
e6b86808f517f2f3379855fcca4ba51ebb5d580e0fa708
!
interface ip 103
        ip ospf enable
        ip ospf message-digest-key 1
!
interface ip 104
        ip ospf enable
        ip ospf message-digest-key 1
!
ipv6 router ospf
        router-id 2.2.2.2
        enable
!
        area 0 area-id 0.0.0.0
        area 0 stability-interval 40
        area 0 default-metric 1
        area 0 default-metric type 1
        area 0 translation-role candidate
        area 0 type transit
```

```
        area 0 enable
!
        redistribute connected export 10 1
!
interface ip 113
        ipv6 ospf area 0
        ipv6 ospf retransmit-interval 5
        ipv6 ospf transmit-delay 1
        ipv6 ospf priority 1
        ipv6 ospf hello-interval 10
        ipv6 ospf dead-interval 40
        ipv6 ospf cost 1
        ipv6 ospf enable
!
interface ip 114
        ipv6 ospf area 0
        ipv6 ospf retransmit-interval 5
        ipv6 ospf transmit-delay 1
        ipv6 ospf priority 1
        ipv6 ospf hello-interval 10
        ipv6 ospf dead-interval 40
        ipv6 ospf cost 1
        ipv6 ospf enable
!
end
```

ACC-3: Access switch (Virtual Fabric 10G Switch Module stack)

Example A-5 shows the final configuration of the ACC-3 access switch.

Example A-5 Final configuration of the ACC-3 access switch

```
version "6.8.0.1"
switch-type "BNT Virtual Fabric 10Gb Switch Module for IBM BladeCenter, Stack"
stack backup 2
stack switch-number 1 universal-unic-id 05e9050bcd92450f903d7e60c581e4a4
stack switch-number 1 bay 7
stack switch-number 2 universal-unic-id 05e9050bcd92450f903d7e60c581e4a4
stack switch-number 2 bay 9
!
!
ssh scp-enable
ssh enable
!
!
hostname "ACC-3"
system idle 60
!
!
interface port 1:14
        name "SRV-3"
        pvid 30
        exit
!
```

```
interface port 1:17
        name "AGG1-ACC3"
        pvid 30
        exit
!
interface port 1:18
        name "AGG2-ACC3"
        pvid 30
        exit
!
interface port 2:14
        name "SRV-3"
        pvid 30
        exit
!
interface port 2:17
        name "AGG1-ACC3"
        pvid 30
        exit
!
interface port 2:18
        name "AGG2-ACC3"
        pvid 30
        exit
!
vlan 1
        member 1:1-1:13,1:19-1:27
        no member 1:14,1:17-1:18
        member 2:1-2:13,2:19-2:27
        no member 2:14,2:17-2:18
        member 3:1-3:14,3:17-3:27
        member 4:1-4:14,4:17-4:27
        member 5:1-5:14,5:17-5:27
        member 6:1-6:14,6:17-6:27
        member 7:1-7:14,7:17-7:27
        member 8:1-8:14,8:17-8:27
!
vlan 30
        enable
        name "SRV-3"
        member 1:14,1:17-1:18
        member 2:14,2:17-2:18
!
vlan 4095
        member 1:1-1:3,1:5-1:13,1:15-1:16
        no member 1:4,1:14
        member 2:1-2:3,2:5-2:13,2:15-2:16
        no member 2:4,2:14
        member 3:1-3:16
        member 4:1-4:16
        member 5:1-5:16
        member 6:1-6:16
        member 7:1-7:16
        member 8:1-8:16
!
```

```
portchannel 1 port 1:17
portchannel 1 port 2:17
portchannel 1 enable
!
portchannel 2 port 1:18
portchannel 2 port 2:18
portchannel 2 enable
!
!
spanning-tree stp 30 vlan 30

!
snmp-server name "ACC-3"
!
!
!
!
!
!
!
!
lldp enable
!
!
!
!
end
```

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this book.

Locating the web material

The web material associated with this book is available in softcopy on the Internet from the IBM Redbooks web server. Point your web browser at:

ftp://www.redbooks.ibm.com/redbooks/SG247960

Alternatively, you can go to the IBM Redbooks website at:

ibm.com/redbooks

Select the **Additional materials** and open the directory that corresponds with the IBM Redbooks form number, SG247960.

IBM Redbooks

The following IBM Redbooks publications provide additional information about the topic in this document. Note that some publications referenced in this list might be available in softcopy only.

- *IBM BladeCenter Products and Technology*, SG24-7523
- *BNT 1/10Gb Uplink Ethernet Switch Module for IBM BladeCenter*, TIPS0705
- *BNT Virtual Fabric 10Gb Switch Module for IBM BladeCenter*, TIPS0708
- *IBM System Networking RackSwitch G8052*, TIPS0813
- *IBM System Networking RackSwitch G8124*, TIPS0787
- *IBM System Networking RackSwitch G8264/G8264T*, TIPS0815

You can search for, view, download or order these documents and other Redbooks, Redpapers, Web Docs, draft and additional materials, at the following website:

ibm.com/redbooks

Other publications

These publications are also relevant as further information sources:

- *IBM 1/10Gb Uplink Ethernet Switch Module for IBM BladeCenter Application Guide*:

  http://www-947.ibm.com/systems/support/supportsite.wss/docdisplay?brandind=5000008&lndocid=MIGR-5076214

- *IBM 1/10Gb Uplink Ethernet Switch Module for IBM BladeCenter BBI Quick Guide*:

  http://www-947.ibm.com/systems/support/supportsite.wss/docdisplay?brandind=5000008&lndocid=MIGR-5076219

© Copyright IBM Corp. 2012. All rights reserved.

- *IBM 1/10Gb Uplink Ethernet Switch Module for IBM BladeCenter Command Reference*:

  http://www-947.ibm.com/systems/support/supportsite.wss/docdisplay?brandind=5000008&lndocid=MIGR-5076525
- *IBM 1/10Gb Uplink Ethernet Switch Module for IBM BladeCenter Installation Guide*:

  ftp://ftp.software.ibm.com/systems/support/system_x_pdf/dw1gymst.pdf
- *IBM 1/10Gb Uplink Ethernet Switch Module for IBM BladeCenter ISCLI Reference*:

  http://www-947.ibm.com/systems/support/supportsite.wss/docdisplay?brandind=5000008&lndocid=MIGR-5076215
- *IBM BladeCenter H Installation and Users Guide*:

  http://publib.boulder.ibm.com/infocenter/bladectr/documentation/topic/com.ibm.bladecenter.8852.doc/bc_8852_iug.html
- *IBM BladeCenter H Trouble Shooting*:

  http://publib.boulder.ibm.com/infocenter/bladectr/documentation/topic/com.ibm.bladecenter.8852.doc/bc_8852_pdsg.html
- *IBM BladeCenter HT Installation and Users Guide*:

  http://publib.boulder.ibm.com/infocenter/bladectr/documentation/topic/com.ibm.bladecenter.8750.doc/bc_8750_iug.html
- *IBM BladeCenter HT Trouble Shooting*:

  http://publib.boulder.ibm.com/infocenter/bladectr/documentation/topic/com.ibm.bladecenter.8750.doc/bc_8750_pdsg.html
- *IBM RackSwitch G8052 Blade OS Application Guide*:

  http://www-01.ibm.com/support/docview.wss?uid=isg3T7000353
- *IBM RackSwitch G8052 Browser-Based Interface Quick Guide*:

  http://www-01.ibm.com/support/docview.wss?uid=isg3T7000348
- *IBM RackSwitch G8052 ISCLI Command Reference*:

  http://www-01.ibm.com/support/docview.wss?uid=isg3T7000344
- *IBM RackSwitch G8052 Installation Guide*:

  http://www.bladenetwork.net/userfiles/file/G8052_install.pdf
- *IBM RackSwitch G8052 Menu-Based CLI Reference Guide*:

  http://www-01.ibm.com/support/docview.wss?uid=isg3T7000347
- *IBM RackSwitch G8124/G8124E Browser-Based Interface Quick Guide*:

  http://www-01.ibm.com/support/docview.wss?uid=isg3T7000389
- *IBM RackSwitch G8124/G8124E ISCLI Command Reference*:

  http://www-01.ibm.com/support/docview.wss?uid=isg3T7000390
- *IBM RackSwitch G8124/G8124E Menu-Based CLI Reference Guide*:

  http://www-01.ibm.com/support/docview.wss?uid=isg3T7000391
- *IBM RackSwitch G8124 Blade OS Application Guide*:

  http://www-01.ibm.com/support/docview.wss?uid=isg3T7000388
- *IBM RackSwitch G8124 Installation Guide*:

  https://www-304.ibm.com/support/docview.wss?uid=isg3T7000299&aid=1

- *IBM RackSwitch G8264 Blade OS Application Guide*:

  http://www-01.ibm.com/support/docview.wss?uid=isg3T7000326
- *IBM RackSwitch G8264 Browser-Based Interface Quick Guide*:

  http://www-01.ibm.com/support/docview.wss?uid=isg3T7000342
- *IBM RackSwitch G8264 Installation Guide*:

  http://www.bladenetwork.net/userfiles/file/G8264_install.pdf
- *IBM RackSwitch G8264 ISCLI Command Reference*:

  http://www-01.ibm.com/support/docview.wss?uid=isg3T7000329
- *IBM RackSwitch G8264 Menu-Based CLI Reference Guide*:

  http://www-01.ibm.com/support/docview.wss?uid=isg3T7000328
- *IBM SNEM 6.1 Release Notes Changes*:

  http://www-01.ibm.com/support/docview.wss?uid=isg3T7000474&aid=1
- *IBM SNEM 6.1 Solution Getting Started Guide*:

  http://www-01.ibm.com/support/docview.wss?uid=isg3T7000471&aid=1
- *IBM SNEM 6.1 User Guide*:

  http://www-01.ibm.com/support/docview.wss?uid=isg3T7000473&aid=1
- *IBM System Networking Element Manager Solution Device Support List (6.1)*:

  http://www-01.ibm.com/support/docview.wss?uid=isg3T7000474
- *IBM Virtual Fabric 10Gb Switch Module for IBM BladeCenter Application Guide*:

  http://www-01.ibm.com/support/docview.wss?uid=isg3T7000496&aid=2
- *IBM Virtual Fabric 10Gb Switch Module for IBM BladeCenter BBI Quick Guide*:

  http://download.boulder.ibm.com/ibmdl/pub/systems/support/system_x_pdf/bmd00192.pdf
- *IBM Virtual Fabric 10Gb Switch Module for IBM BladeCenter Command Reference*:

  http://download.boulder.ibm.com/ibmdl/pub/systems/support/system_x_pdf/bmd00190.pdf
- *IBM Virtual Fabric 10Gb Switch Module for IBM BladeCenter Installation Guide*:

  http://download.boulder.ibm.com/ibmdl/pub/systems/support/system_x_pdf/46m1525.pdf
- *IBM Virtual Fabric 10Gb Switch Module for IBM BladeCenter ISCLI Reference*:

  http://download.boulder.ibm.com/ibmdl/pub/systems/support/system_x_pdf/bmd00191.pdf
- *Quick Start Guide for Installing and Running KVM*:

  http://publib.boulder.ibm.com/infocenter/lnxinfo/v3r0m0/topic/liaai/kvminstall/kvminstall_pdf.pdf

Online resources

These websites are also relevant as further information sources:

- IBM 1/10Gb Uplink Ethernet Switch Module Announcement Letter:

  http://www.ibm.com/common/ssi/rep_ca/5/872/ENUSAG08-0365/ENUSAG080365.PDF

- IBM BladeCenter H Announcement Letter:

  http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?infotype=an&subtype=ca&appname=gpateam&supplier=897&letternum=ENUS109-438

- IBM BladeCenter HT Announcement Letter:

  http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?infotype=AN&subtype=CA&htmlfid=897/ENUS110-209&appname=USN

- IBM RackSwitch G8052 Announcement Letter:

  http://www.ibm.com/common/ssi/cgi-bin/ssialias?infotype=AN&subtype=CA&appname=gpateam&supplier=872&letternum=ENUSAG11-0005&pdf=yes

- IBM RackSwitch G8124 Announcement Letter:

  http://www.ibm.com/common/ssi/cgi-bin/ssialias?infotype=AN&subtype=CA&appname=gpateam&supplier=899&letternum=ENUSLG11-0096&pdf=yes

- IBM RackSwitch G8264 Announcement Letter:

  http://www.ibm.com/common/ssi/cgi-bin/ssialias?infotype=AN&subtype=CA&appname=gpateam&supplier=872&letternum=ENUSAG11-0005&pdf=yes

- IBM System Networking Options:

  http://www-03.ibm.com/systems/networking/options/

- IBM Virtual Fabric 10Gb Switch Module Announcement Letter:

  http://www.ibm.com/common/ssi/rep_ca/5/872/ENUSAG09-0245/ENUSAG09-0245.PDF

- Storage Networking Industry Association (SNIA):

  http://www.snia.org

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

Index

Numerics

10 Gb uplink ports 7
10Gb SFP+ 48
10GbE SFP+ ports 8
1Gb SFP+ 48
802.1p configuration 183
802.1p priority value 271
802.1Q VLAN tagging 167, 262
802.1Q VLAN tags 53
802.1x Port-based Network Access Control 95

A

Access Control Lists (ACLs) 97, 101
Access Layer 111
Access layer switches 119
Accounting 95
ACL actions 99
ACL configuration 171
ACL Metering 102
ACL-based mirroring 10, 18, 25, 75
ACLs 75–76, 97
Active MultiPath (AMP) 14, 17
Active Multi-Path Protocol (AMP) 241
Active MultiPath Protocol (AMP) 86
Active-Active Redundancy 84
Address Resolution Protocol (ARP) 83
adjacencies 64, 208, 215
Admin key 78, 175
Administrator 139
Advanced Management Module (AMM) 142
Aggregation layer 108
AMM 45
AMM interface 127, 146
AMP 86
AMP links 87
anycast 72, 193
anycast address 195
ARAP 94
Area Border Router (ABR) 64
area ID 205, 212
area index 205, 212
ARP broadcast packets 184
ARP table 80
ARP unicast reply packets 184
ASIC 8, 60
asnum 244, 253
Assured Forwarding (AF) 103
Attached Switch Number (asnum) 244
Authentication Server 96
Authenticator 96
Auto-Fallback 81
auto-negotiation 157, 257
Autonomous System (AS) 62
Autonomous System Boundary Router (ASBR) 64

B

Backup Designated Router (BDR) 64
Backup Switch 243
bay number 253
BBI 124–125
BCM rate control 241
BDR 64
BGP 7, 14, 17, 21, 25, 30, 66, 70, 186, 193, 219, 240
BGP Aggregation 222
BGP configuration mode 220
BGP global configuration 220
BGP Peer Configuration 220
BGP peers 221
BGP Redistribution 221
Blade Center architecture 108
BLADE Harmony Manager 123
BladeCenter Advance Management Module interface 124
BNT Virtual Fabric 10 Gb Switch Module 27
  See also Ethernet switch modules
  cables 32
  DAC cables 48
  direct-attached cables 28
  features 29
  IEEE standards 32
  Layer 3 functions 30
  management 31
  performance 29
  ports 29
  QoS 30
  redundancy 30
  scalability 29
  security 30
  SFP+ transceivers 28
  shipgroup 28
  transceivers 28, 48
  VLANs 30
boot image version 291
Boot Management Menu 291
Bootstrap Protocol (BOOTP) 193
Border Gateway Protocol 14, 193
Border Gateway Protocol (BGP) 7, 66, 193, 219, 241
border routers 66
BPDU Guard 120
BPDU guard 17, 25
BPDUs 59, 86
bridge priority 180
Bridge Protocol Data Units (BPDUs) 59
Broadcast Storms 100
Browser-Based interface (BBI) 125

C

CEE 22, 28
CFF-h 47

© Copyright IBM Corp. 2012. All rights reserved.

CHAP 94
chassis UUID 253
CIO-v 46
Cisco STP packets 184
cKVM 39
class of service 138
Class of Service (COS) 105
Class Selector (CS) 104
CLI 129, 286
CLI commands 91
Client 96
Combination I/O Vertical (CIO-v) 41
Combo Form Factor Horizontal (CFF-h) 42
Community 90
Community VLAN 89
community VLAN 264
compact flash (CF) 44
concurrent KVM (cKVM) 39
Configuration blocks 286
Configuration files 286
Configure authentication 206
Configure hashing 172, 266
Configured Switch Number (csnum) 244
Configuring date and time 2
Console 125
control plane (CP) 184
Control Plane Protection 184
Converge Enhanced Ethernet (CEE) 241
Converged Enhanced Ethernet (CEE) 14, 27
Converged Enhanced Ethernet (CEE) 17, 24
Converged fabric 14, 22
COS 105, 138
CRC trailer 52
csnum 244, 253

D

DAC cables 20
Data bits 127
Data Center Bridge Exchange protocol (DCBX) 17, 24
Data Center Bridging (DCB) 50
Date & time 131
date and time 124
daughter card 41
DCBX 31
Default baud rate 127
Default Gateway 135
Default gateway 188
default gateway 136, 188
Default Password 138
default password 128
Designated Router (DR) 64, 70
destination IP 172
Destination IPv4 (DIP) 83
destination MAC 172
Destination port 191
DHCP 18, 25
DHCP packets 184
DHCP Relay 35
DHCPv6 73, 193
Differentiated Services (DS) 103
DiffServ Code Point (DSCP) 183
DiffServ Code Point value 271
DiffServ Code Points (DSCP) 102
Dijkstra's algorithm 65
DIP 267
dip 191
DIP (destination IP only) 172
dipsip 190
Distance Vector Protocol 61
DMAC 267
DMAC (destination MAC only) 172
Double-Wide HX5 46
dport 191
DR 64, 70
DSCP 102, 183, 272
DSCP configuration 183
duplex 131, 161, 171
duplex mode 157, 257
Dynamic Host Configuration Protocol (DHCP) 30
Dynamic Host Control Protocol 193
dynamic LACP trunk groups 77, 171, 265
dynamic routing protocols 155
Dynamic trunk 176
Dynamic Trunks 175, 268

E

EAPoL 96
EAPoL Authentication 96
eBGP 66, 219
ECMP 61
ECMP gateway 190
ECMP hash setting 190
ECMP health-check 191
ECMP route hasing 190
ECMP static route 192
ECMP Static Routes 61
ECMP static routes 190
ECN 185
Egress packets 105
election priority 225, 276
Enhanced Transmission Selection (ETS) 17, 24, 31
Equal-Cost Multi-Path (ECMP) 61, 190
Ethernet switch modules
See also BNT Virtual Fabric 10 Gb Switch Module
Event List Report 305
Expedited Forwarding (EF) 103
Explicit Congestion Notification (ECN) 185
Extensible Authentication Protocol 95
Extensible Authentication Protocol (EAP) 96
External BGP (eBGP) 66
external BGP (eBGP) 219
external routing 65

F

factory defaults 289
failed firmware upgrade 293
Failover Limit 81
Failover Manual Monitor Control 231
Failover Manual Monitor Port 231

Failover Methods 84
failover mode 193
Failover Trigger 232
failover trigger 233
Failover Trigger limit 232
failover trigger status 234–235
Fast Uplink Convergence 80
FastLeave 68
FCoE 22, 28, 31
FCoE Initialization Protocol (FIP) 31
FDB 80
FDB Flush 88
FDB Update 80
Fibre Channel over Ethernet (FCoE) 27, 50, 241
Fibre Channel over Ethernet (FCoE) 17
FIP Snooping 17, 24
firmware 292–293
Firmware files 290
Firmware Management 290
Firmware upgrade 149
Flow control 127
flow control 18, 131, 157, 161, 171, 257
Forward Delay 80
Forwarding Database (FDB) 80
full duplex 22
full-duplex 157, 257
Full-duplex Flow Control 32, 36

G

G8052 6
G8052F 6
G8052R 6
G8124 12
G8124DC 13, 15
G8124E-F 13, 15
G8124E-R 13, 15
G8124F 13, 15
G8124R 13, 15
G8264 20
G8264F 21
G8264R 21
gateway 130
Gigabit Ethernet connector 132
Gigabit Ethernet interface 157
Gigabit Ethernet port flow parameters 132
global commands 180
Global Configuration mode 130
global hash parameters 174, 267
global LACP parameters 176
global routing table 216

H

Health Checks 87
Health Status 305
High availability 10
High Speed Switch (HSS) 144
host nodes 194
Hot Link 7, 14, 22
Hot Links 3, 10, 17, 25, 79, 222, 240, 272, 274
HTTPS 24

I

iBGP 66, 219
IBM 1/10 Uplink Ethernet Switch Module 33
IBM 1000BASE-SX SFP Transceiver 48
IBM 1000BASE-T (RJ45) SFP Transceiver 48
IBM 10GBase-SR 10GbE 850 nm Fiber SFP+ Transceiver 48
IBM BladeCenter E 38
IBM BladeCenter H 38
IBM BladeCenter HT 43
IBM BladeCenter Switches 27
IBM Networking Operating System 156
IBM Networking Operating System 6.8 163
IBM Networking OS 129, 155
IBM Networking OS 6.8 163, 260
IBM Networking OS Command Line Interface (CLI) 125
IBM Networking OS implementation 192
IBM Networking OS web interface dashboard 147
IBM RackSwitch switches 157
IBM SFP+ Transceiver 48
IBM System Networking 1/10Gb Uplink Ethernet Switch Module 4
IBM System Networking 10Gb ethernet switches 51
IBM System Networking Element Manager 10, 28, 123
IBM System Networking Element Manager (SNEM) 2, 150, 303
IBM System Networking RackSwitch G8052 4
IBM System Networking RackSwitch G8124 4, 12
IBM System Networking RackSwitch G8264 4
IBM System Networking RackSwitches 48
IBM System Networking Virtual Fabric 10GB Network switch 4
IBM Systems Director 124
IBM Tivoli Network Manager version 1 for SNEM 151
IBM Virtual Fabric 13, 20
IBM Virtual Fabric 10Gb module 124
IBM Virtual Fabric 10Gb Switch Module 27–28, 48, 239
ICMP 151, 284
ICMP packets 184
ICMP test 219
IEEE 802.1p standard 105
IEEE 802.1x packets 184
IGMP 25, 29, 67, 69, 193
IGMP Entries 68
IGMP Filtering 69
IGMP filtering 34
IGMP Membership Queries 67
IGMP packets 184
IGMP Proxy 69
IGMP Querier 69
IGMP Querier election 67
IGMP Relay 69, 241
IGMP snooping 34
IGMP v3 Snooping 7, 68
IGMPv3 68, 241
Independent Stacks 243
Industry Standard CLI 129
InfiniBand 39, 45

Intel Connects Optical Cable 48
interface status 234, 236
internal BGP (iBGP) 66, 219
Internal Router (IR) 64
internal routing 65
Internet Group Management Protocol (IGMP) 67, 193
Internet Protocol version 6 (IPv6) 71
inter-switch connection 111
IP address 130, 144
IP Addressing 113
IP Configuration 134
IP filtering 35
IP forwarding 30, 35, 136
IP interface 186–187
IP multicast 67
IP Routing 60, 136
IP routing 155
IP subnets 60, 136
IPv4 108, 121, 134, 186, 192
IPv4 packets 184
IPv4 static routes 189
IPv6 18, 25, 71, 108, 121, 186, 192, 240–241
IPv6 ACLs 97
IPv6 address 71
IPv6 address types 72
IPv6 interface 193
IPv6 Neighbor Discovery packets 184
ISCLI 292
Isolated 90
Isolated VLAN 89
isolated VLAN 264

J

Jumbo frame 52
jumbo frames 18, 34

L

LACP 17, 24, 78, 82, 175
LACP admin key 177
LACP adminkey 80
LACP configuration 177
LACP mode 176
LACP modes 78, 175
LACP negotiation 176
LACP parameters 177
LACP port priority 176
LACP ports 176
LACP priority value 176
LACP trunk group 78, 177
LACPDU 78, 175
LAG ID 78, 175
Layer 1 155, 159, 240, 255
Layer 1 architecture 109, 117
Layer 1 configuration 156
Layer 2 155, 163, 240, 260
Layer 2 architecture 111, 119
Layer 2 Failover 222, 230, 272
Layer 2 failover 233, 240
Layer 3 155, 186
Layer 3 architecture 113, 120
Layer 4 TCP traffic 191
Layer 4 UDP traffic 191
LDAP 9, 95, 141, 193
LDAP Authentication 95
LDAP authentication 141
LED Status 26, 306
LED status 19
Lightweight Directory Access Protocol 95
Lightweight Directory Access Protocol (LDAP) 30
Link Aggregation Control Protocol 32, 36, 78
Link Aggregation Control Protocol (LACP) 5, 29, 50, 81, 175
Link Aggregation Identifier (LAG ID) 78, 175
Link Layer Detection Protocol (LLDP) 241
Link State Database (LSDB) 62, 64
Link-State Advertisement (LSA) 65
Linux Kernal-based Virtual Machine (KVM) 152
LLC 56
LLDP 10, 18, 25
LLDP packets 184
Logging console 297
Logging destinations 297
Logical Link Control 32, 36, 56
Loopback Interfaces 241
Loopback interfaces 115
Low Latency 14
LSA 65
LSDB 62, 64

M

MAC Address 95
MAC address 58, 183
MAC address hashing 173
MAC address notification 241
MAC table 18
management interface 187
Management IP addresses 188
management ports 130
Master Failover 244
Master Recovery 244
Master Switch 242
Master VRRP Router 83
Maximum Transmission Unit (MTU) 52
MD5 cryptographic authentication 206
MD5 key ID 207
Mean time between failures (MTBF) 16
Media access control (MAC) 29
Member switch 254
Member switches 243
Menu-Based CLI 292
Metering 102
Molex Direct Attach Copper SFP+ Cable 48
Monitoring Trunk Links 81
Mrouter 68–69
MSTP 32, 59, 179, 241, 269
MTU 52
Multicast 10
Multicast addresses 72
Multicast Router (Mrouter) 68

Multiple Spanning Tree 24
Multiple Spanning Tree Protocol (MSTP) 59
Multiple STP (MSTP) 30, 34, 36
Myrinet 45

N

negotiation mode 131
Neighbor Advertisements 73
Neighbor Discovery 73
Neighbor Discovery protocol (ND) 73
Neighbor Solicitations 73
neighbors 64
Netboot 18
Network Access Server (NAS) 94
Network Adapter Teaming 273
Network Equipment Provider (NEP) 43
Network Time Protocol (NTP) 35
network topology 107
NIC Teaming 80
NIC teaming 5, 193
No Backup 244
Non-blocking architecture 34
Not-So-Stubby-Area (NSSA) 63
NSSA 63
NTP 10

O

Open Shortest Path First 14
Open Shortest Path First (OSPF) 7, 62, 115
Open Shortest Path First protocol (OSPF) 30
Operator 139
OSI Layer 155
OSPF 7, 14, 17, 21, 25, 62, 64, 70, 113, 186, 189–190, 202, 240–241
OSPF area 204
OSPF database 209
OSPF domain 205, 212
OSPF integration 190
OSPF packets 184
OSPF protocol 202
OSPF3 Packets 184
OSPFv2 115, 204, 217
OSPFv3 30, 115, 202, 211, 217, 241
OSPFv3 areas 212
OSPFv3 configuration 214
Oversubscription 29

P

PAP 94
Parity 127
passive-interface 115
password recovery 290
Per-Hop Behavior (PHB) 103
Per-VLAN Rapid Spanning Tree (PVRST) 179, 269
Per-VLAN Rapid Spanning Tree Protocol (PVRSTP) 112, 120
Per-VLAN Rapid Spanning-Tree Protocol (PVRST) 59
PHB 103
Physical Layer 109
PIM 69
PIM Dense Mode 22
PIM Dense Mode (PIM-DM) 70
PIM packets 184
PIM Sparse Mode 22, 25
PIM Sparse Mode (PIM-SM) 70
PIM-DM 70
PIM-SM 70
Ping 307
Ping request 145
Port flood blocking 242
Port link 257
Port link configuration 157
Port link LED 306
Port Mirroring 76
Port mirroring 10, 18, 25, 74
port mode 158
Port settings 256
port settings 156
port trunk groups 170
Port VLAN ID Numbers 164, 261
Port VLAN identifier (PVID) 53
port-based authentication 32, 36
Port-based VLANs 17
PortFast 5
Ports 163, 260
ports parameters 180
Power Modules 40
Preemption 80
preemption 225, 277
Primary VLAN 264
Priority Based Flow Control (PFC) 24
Priority-Based Flow Control (PFC) 31
Priority-based Flow Control (PFC) 17
Private VLAN 170
Private VLANs 89, 164, 261, 264
Promiscuous 90
Protected Mode 91
Protocol Independent Multicast (PIM) 67, 69, 193
Protocol-based VLAN 169
Protocol-based VLANs 164, 242
Protocol-based VLANs (PVLANs) 56
PVID 52–53, 165, 167, 170
PVID assignment 166
PVLAN 56–57
PVRST 59, 179–180, 269
PVRST+ 17, 24, 32, 50
PVRSTP 112, 120

Q

QLogic Virtual Fabric Extension Module 28
QoS 3, 97, 100, 155, 183, 193, 240, 271
QoS 802.1p 105
QoS Levels 104
QSFP+ 20
QSFP+ 40GbE mode 110
QSFP+ port 158
QSFP+ ports 22
Quality of Service 163, 260

Quality of service 10, 17
Quality of Service (QoS) 35, 100, 271
Querier 67
query-response-interval 68

R

RackSwitch G8052 6–7
RackSwitch G8124 12, 155
RackSwitch G8264 20
RADIUS 9, 16, 24, 91–92, 139, 193
Radius 30, 35
Radius authentication 139
RADIUS server 139
Rapid Spanning Tree 5, 24
Rapid Spanning Tree Protocol 50
Rapid Spanning-Tree Protocol (RSTP) 59
Rapid STP (RSTP) 30, 34, 36
RAS 92
Redbooks website 333
  Contact us xiv
Redirect messages 73
Redundant midplane 44
Reference architectures 108
Re-Maring 102
Re-Marking 102
Remote Access Server (RAS) 92
Remote Authentication Dial-in User Service 92
Remote Monitoring (RMON) 76, 301
Rendezvous Point (RP) 70
Reset 159
Restore Factory Defaults 145
ring topology 242
RIP 25, 35, 61, 70, 242
RIP packets 184
RIPng 193
RIPv1 61–62
RIPv2 61–62
RMON 10, 18, 25, 76, 301
Route maps 242
route redistribution 207, 213
Router Advertisement protocol 201
Router Advertisements 73, 194–195
Router Advertisements protocol 198
Router ID 211
Router IDs 242
Router Information Protocol (RIP) 30
router nodes 194
Router Solicitations 73
Routing 136
Routing Information Protocol 193
Routing Information Protocol (RIP) 61, 193
Routing protocol support 35
Routing protocols 10
routing table 105
RP 70
RS-232 serial console port 8
RS-232 serial port 34
RSTP 32, 50, 59, 179, 269

S

Sample Rate 76
saved switch configuration 288
SCP 17, 24
Secondary Backup 244
Secure Copy 90
Secure Copy (SCP) 90
Secure Shell 90
Secure Shell (SSH) 90, 125
Security 2, 124
Serial over LAN (SOL) 31
Setting up the IP address 2
Setup utility 137
setup utility 131
sFLOW 193
sFlow 75
sFlow Network Sampling 76
sFlow port monitoring 242
sFlow Statistical Counters 75
SFP+ 6, 15
SFP+ copper direct-attach cables (DAC) 34
SFP+ direct-attach copper (DAC) 29
SFP+ modules 34
SFP+ ports 20, 22
SFP+ transceivers 6, 13, 29
Shortest Path First 65
Simple Network Management Protocol 31, 35
Simple Network Management Protocol (SNMP) 297
simple static route 192
SIP 267
sip 190–191
SIP (source IP only) 172
SMAC 267
SMAC (source MAC only) 172
SNAP 56
SNEM 150
SNEM Alerts Report 305
SNMP 18, 25, 31, 35, 124, 151, 287, 297
SNMP agent 295
SNMP trap 193
SNMP traps 295, 299
SNMPv2 Trap 299
SNMPv3 298
SNMPv3 Trap 300
source IP 172
source IP address (SIP) 190
source MAC 172
Source port 191
Source-Specific Multicast (SSM) 68
Spanning Tree 9, 17, 171
Spanning Tree configuration 181
Spanning Tree Group (STG) 265
Spanning Tree Group membership 134
Spanning Tree Groups 179, 269
Spanning Tree Groups (STGs) 179
Spanning Tree Protocol 3, 79, 131, 179, 240
Spanning Tree Protocol (STP) 29, 36, 81, 112
Spanning Tree Protocol packets 184
Spanning Tree Protocols 163, 260
Spanning Tree protocols 155

Spanning-Tree mode 120
Spanning-Tree Protocol (STP) 58
Speed 131
speed 161, 171, 257
sport 191
SSH 17, 24, 91, 129
SSM 68
stack 88
stack links 242
Stack Member Identification 244
stack members 249
Stack Membership 242
stack status 248
Stacking 3, 88, 272
stacking 240
stacking configuration 247
stacking links 246
stacking membership mode 246
stacking VLAN 246
Stateful address configuration 73
Stateless address configuration 73
stateless auto-configuration 194
Static MAC address adding 242
Static multicast 242
static route health check 192
Static routes 61
static routes 70, 189
static trunk groups 77, 265
static trunk groups (portchannel) 171
Static Trunks 17, 24, 82, 171
STG 179
STG Assignment 269
Stop bits 127
Storage Networking Industry Association (SNIA) 49
storm control 171
Storm-Control Filters 100
STP 32, 34, 58, 82, 86, 120
STP bridge priorities 121
STP root bridge 121
STP/PVST+ 179
Stub area 63
Subnetwork Access Protocol 56
Supplicant 96
Support for Serial over LAN (SOL) 36
switch dump 288
switch image 287
Switch Version Report 305
Syslog 296
Syslog List Report 305
System ID 78, 175
System LED 306
systems logs 295

T

TACACS 24, 91, 140–141
TACACS+ 9, 16, 35, 91, 193
TACACS+ Authentication 94
TACACS+ Command Authorization 140
Tagged frame 53
Tagged member 53
Tagged Packets 32, 36
Tagged packets 105
Tagged VLAN 32, 36
tagging 3, 167, 240
TCP flag 99
TCP port number 141–142
tcpl4 191
Teaming Adapter 197
Telnet 91, 125
Telnet access 146
Temperature Sensor Warning 307
Terminal Access Controller Access-Control System Plus (TACACS+) 30
terminal client configuration 127
Terminal connection 2, 124
TFTP 18
three-tier Data Center design 108
throughput 7, 20
timers 180, 225
Time-To-Live (TTL) 308
Time-to-live (TTL) 221
Tivoli Application Dependency Discovery Manager (TADDM) 152
Tivoli Netcool Configuration Manager 152
Tivoli Netcool/OMNIbus 152
Top of Rack (TOR) 2, 123
Top-of-Rack architecture 108
Top-of-Rack switches 163
TOR switch 127
Traceroute 308
tracking 227, 278
traffic statistics 162, 259
Transceiver Information Report 305
transceivers 159
Transit Area 63
troubleshooting 306
trunk failover 22
trunk group 266
trunk group parameters 175, 268
trunk group status 174
Trunk Hash algorithm 17
trunk hash algorithm 77
Trunk Links 241
Trunking 3, 9, 77, 163, 222, 236, 260, 272–273
trunking 155, 240
Trunks 113, 241
TTL 221

U

UDLD packets 184
UDP 62
UDP port 140
udpl4 191
unicast 193
Unicast Addreses 72
unicast IPv6 address 195
Uni-Directional Link Detection (UDLD) 242
Untagged frame 53
Untagged member 53
Untagged packets 105

untagged port 169
Uplink Failure Detection 81
Uplink failure detection 10, 17
Uplink Failure Detection (UFD) 7, 14
USB port 8
User 138
User Account 138
user accounts 137
User Datagram Protocol (UDP) 62

V

VASA 179, 269
VFSM 96, 239, 274
VID 53
Virtual Fabric 14, 21
Virtual Fabric 10Gb Switch Module 239
Virtual Interface Router 83
virtual link 170
Virtual Link Aggregation Groups (VLAGs) 79
Virtual Local Area Networks (VLANs) 52, 60
Virtual local area networks (VLANs) 111
Virtual NICs 242
virtual NICs 13
virtual NICs (vNICs) 20
Virtual Router 82, 223, 275
virtual router group 84
virtual router identifier (VRID) 82
Virtual Router IP address 275
Virtual Router MAC Address 83
Virtual Router Redundancy Protocol (VRRP) 30, 34, 193, 242
Virtual Router Redundancy Protocol(VRRP) 222
Virtual Router Redundancy support (VRRP) 10, 17, 25
Virtual Router tracking 229
VLAG packets 184
VLAGs 79, 86
VLAN 69, 131, 135, 186
VLAN 4095 187
VLAN Automatic STG Assignment (VASA) 179, 269
VLAN configuration 133, 165
VLAN ID 170
VLAN identifier (VID) 53
VLAN Maps 97
VLAN member ports 167
VLAN number 135
VLAN port assignment 165
VLAN tag 53
VLAN Tagging 164, 167, 261–262
VLAN tagging 52, 131
VLANs 3, 24, 52, 111, 120, 155, 163–164, 240–241, 260–261, 263
VLAN-tagging adapter 167
VM aware Networking 14
VM Data Center Report 305
VMAP 100
VMap 75
VM-Aware Networking 21
VMready 10, 18, 21, 25, 31, 193
VMready VM Report 305
VMready® 7
VMware ESX/ESXi 152
vNICs 13
Voice over IP (VoIP) 28
VRID 82, 224
VRRP 3, 5, 7, 14, 22, 82, 113, 155, 180, 223, 240, 272, 274
VRRP configuration 121
VRRP packets 184
VRRP priorities 121
VRRP router 82
VRRP statistics 229
VRRP Tracking 227
VRRP tracking priority 229

W

Web access 91, 147
Weighted Random Early Detection (WRED) 185
Weighted Round Robin (WRR) 30, 35
WRED 185
WRED Transmit Queue 185

Implementing IBM System Networking 10Gb Ethernet Switches

(0.5” spine)
0.475”<->0.873”
250 <-> 459 pages

Implementing IBM System Networking 10Gb Ethernet Switches

Introduction to IBM System Networking RackSwitch hardware

Sample network design and implementation

Switch troubleshooting and maintenance

In today's infrastructure, it is common to build networks based on 10 Gb Ethernet technology. The IBM portfolio of 10 Gb systems networking products includes Top-of-Rack switches, and the embedded switches in the IBM BladeCenter family. In 2010, IBM formed the IBM System Networking business (by acquiring BLADE Network Technologies), which is now focused on driving data center networking by using the latest Ethernet technologies.

The main focus of this IBM Redbooks publication is on the IBM System Networking 10Gb Switch Modules, which include both embedded and Top-of-Rack (TOR) models. After reading this book, you can perform basic to advanced configurations of IBM System Networking 10Gb Switch Modules.

In this publication, we introduce the various 10 Gb switch models that are available today and then describe in detail the features that are applicable to these switches.

We then present two architectures that use these 10 Gb switches, which are used throughout this book. These designs are based on preferred practices and the experience of authors of this book. Our intention is to show the configuration of the different features that are available with IBM System Networking 10Gb Switch Modules. We follow the three-tier Data Center design, focusing on the Access and Aggregation Layers, because those layers are the layers that IBM System Networking Switches use.

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

For more information:
ibm.com/redbooks

SG24-7960-00 ISBN 0738436771